Toward the Integration of Behavioral Sensing and Artificial Intelligence

Subigya K. Nepal

Dartmouth College, subigya.k.nepal.gr@dartmouth.edu

Follow this and additional works at: https://digitalcommons.dartmouth.edu/dissertations

Part of the Artificial Intelligence and Robotics Commons, Data Science Commons, Other Computer Sciences Commons, and the Psychology Commons

Recommended Citation
Nepal, Subigya K., "Toward the Integration of Behavioral Sensing and Artificial Intelligence" (2024). Dartmouth College Ph.D Dissertations. 244.
https://digitalcommons.dartmouth.edu/dissertations/244

This Thesis (Ph.D.) is brought to you for free and open access by the Theses and Dissertations at Dartmouth Digital Commons. It has been accepted for inclusion in Dartmouth College Ph.D Dissertations by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
TOWARD THE INTEGRATION OF BEHAVIORAL SENSING AND ARTIFICIAL INTELLIGENCE

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in
Computer Science
by
Subigya K. Nepal
DARTMOUTH COLLEGE
Hanover, New Hampshire
May 2024

Exchanging Committee:

Andrew T. Campbell, Ph.D., Chair

Nicholas C. Jacobson, Ph.D.

Sarah M. Preum, Ph.D.

Mary P. Czerwinski, Ph.D.

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies
Abstract

The integration of behavioral sensing and Artificial Intelligence (AI) has increasingly proven invaluable across various domains, offering profound insights into human behavior, enhancing mental health monitoring, and optimizing workplace productivity. This thesis presents five pivotal studies that employ smartphone, wearable, and laptop-based sensing to explore and push the boundaries of what these technologies can achieve in real-world settings. This body of work explores the innovative and practical applications of AI and behavioral sensing to capture and analyze data for diverse purposes. The first part of the thesis comprises longitudinal studies on behavioral sensing, providing a detailed, long-term view of how significant events, such as the COVID-19 pandemic and professional promotions, impact mental health and productivity among college students and information workers. The College Experience Study spans five years, tracking two cohorts of students through their entire college experience, offering a rare longitudinal perspective on student behavior and mental health during global disruptions. The second study provides a year-long observation of information workers, focusing on physiological and behavioral changes post-promotion, highlighting how career advancements influence personal well-being and professional duty. The second part of the thesis expands the application of behavioral sensing with novel methods and integration with AI technologies. This section introduces a pioneering approach to mental health assessment through the MoodCapture Study, which utilizes passively captured smartphone images to detect depression
symptoms. Additionally, we explore the enhancement of personal productivity and well-being through AI-driven tools, including Large Language Models (LLM) powered personalized productivity agents and the MindScape mobile application for contextual AI-driven journaling. These tools leverage real-time behavioral data to provide adaptive, user-specific support, demonstrating the effectiveness of integrating advanced AI with behavioral sensing in creating responsive and user-centric applications. Overall, our work highlights significant advancements in AI and behavioral sensing, providing practical applications and implications for improving mental health, enhancing educational experiences, and boosting workplace productivity. Each study contributes to a broader understanding of the dynamic interaction between technology and human behavior, laying the groundwork for future innovations that could lead to more efficient, healthy, and productive lifestyles.
Acknowledgments

As I write this, it feels surreal to have finally reached this milestone. Despite moments of doubt, persistence and determination have paid off, and I am grateful to have had the privilege of working with an exceptional team of mentors, colleagues, and friends who have supported me throughout my PhD journey. This team has been instrumental in my growth, and I am thankful for the opportunity to have learned from and collaborated with each of them.

First and foremost, I extend my heartfelt gratitude to my advisor, Andrew T. Campbell, for his guidance, support, and encouragement. Working with Andrew has been a fulfilling experience, challenging at times, yet immensely rewarding. I am also grateful for the mentorship of Mary Czerwinski and Javier Hernandez, as well as the entire Human Understanding and Empathy Group at Microsoft Research, including Jina Suh, Gonzalo Ramos, and Judith Amores. My time with them has been enriching, and I have learned significantly from their expertise.

My colleagues and friends have also played a vital role in my journey. I would like to express special thanks to Shayan Mirjafari, Mikio Obuchi, Weichen Wang, Arvind Pillai, Kizito Masaba, Qijia Shao, Colman Leung, Rob Lewis, Talie Massachi, Orson Xu, and many others. I am grateful for their companionship, collaborations, and discussions, which have not only enhanced my research but also made me a better person.

Reflecting on my path to this point is overwhelming. From my humble begin-
nings in Nepal, where research opportunities were scarce, to attending a prestigious institution like Dartmouth, is a dream realized. I am grateful for the sacrifices and support of my family, who have been my rock throughout this journey. My parents, who valued education above all else, even when resources were limited, deserve special thanks. My brother, who encouraged me to pursue a PhD, has been a constant source of guidance and inspiration. And my partner, Ashmita, has been my rock, best friend, and confidant, providing unwavering support and encouragement throughout this journey.

As I look ahead, the future of my work remains uncertain. However, my hope is that my research contributes to human knowledge, no matter how small, and positively impacts lives. If my efforts can improve even one person’s life, I will consider my work a success, serving a purpose far greater than myself. In closing, I extend my heartfelt thanks to all who have been part of this incredible journey. Your support has not only helped me survive the challenges but thrive amidst them.
## Contents

Abstract ................................................................. ii
Preface ................................................................. iv

1 Introduction ......................................................... 1
  1.1 Overview ...................................................... 1
  1.2 Innovations in Human Behavior Analysis ....................... 3
      1.2.1 Capturing Life Events and Transitions with Longitudinal Behavior Sensing ................................... 4
      1.2.2 Assessing Mental Health via In-the-wild Smartphone Images ...................................................... 6
      1.2.3 Enhancing Personal Productivity and Well-being through LLM and Behavioral Sensing Integration .......... 7
  1.3 Protection of Human Subjects .................................. 9
  1.4 Thesis Contributions .......................................... 11

2 Capturing the College Experience ................................ 15
  2.1 Introduction .................................................... 15
  2.2 Related Work .................................................. 20
  2.3 College Experience Study ...................................... 23
      2.3.1 Study Design ............................................. 23
      2.3.2 Recruitment .............................................. 24
      2.3.3 Demographics ............................................ 25
2.3.4 Sensor Data .......................................................... 25
2.3.5 Ecological Momentary Assessments .............................. 27
2.3.6 Data Completeness, Compliance and Challenges ............ 27
2.4 Four Years of College Life ........................................... 32
2.4.1 COVID Strikes ...................................................... 41
2.5 First vs. Final Term .................................................. 44
2.6 Gender Based Differences .......................................... 47
2.7 Changing Living Conditions ....................................... 50
2.8 Lasting Impact of COVID-19 ...................................... 55
2.8.1 Passively Sensed Behavioral Patterns ........................... 57
2.8.2 Self-reported Mental Health Outcomes ......................... 60
2.9 Discussion ............................................................. 63
2.9.1 Summary of findings ............................................. 63
2.9.2 Implications ....................................................... 65
2.9.3 Ethical Considerations .......................................... 67
2.10 Limitations and Future Work ..................................... 68
2.11 Study Dataset ....................................................... 71
2.12 Conclusion .......................................................... 73

3 Detecting Job Promotion Using Mobile Sensing ................. 75
3.1 Introduction .......................................................... 75
3.2 Related Work ........................................................ 78
3.3 Methodology ......................................................... 81
3.3.1 Study Design ..................................................... 81
3.3.2 Demographic Descriptors ..................................... 82
3.3.3 Promotion Ground Truth ....................................... 82
3.3.4 Mobile Sensing System ........................................ 83
4.4 Results ................................................................. 139
  4.4.1 Image Characteristics ........................................... 140
  4.4.2 Predictive Analysis ............................................. 141
  4.4.3 Ablation Study .................................................. 144
  4.4.4 Machine Learning Feature Importance ....................... 145
  4.4.5 Investigating Bias in Machine Learning ....................... 146
4.5 Ethics ................................................................. 148
4.6 Discussion ........................................................... 151
  4.6.1 Summary of results ............................................. 151
  4.6.2 Implications ..................................................... 153
4.7 Limitations .......................................................... 155
4.8 Conclusion .......................................................... 157
4.9 Acknowledgment ..................................................... 158

5 Personalized Productivity Solutions ................................ 159
  5.1 Introduction ....................................................... 159
  5.2 Related Work ....................................................... 163
  5.3 Methodology ....................................................... 166
    5.3.1 Understanding User Preferences for Productivity Agents ... 167
    5.3.2 Productivity Solutions Design and Evaluation ................ 168
  5.4 Results ............................................................. 177
    5.4.1 Exploring User Preferences for Productivity Agents .......... 177
    5.4.2 Design of the AI Productivity Agent Informed by User Preferences 181
    5.4.3 Productivity Solution Showdown: AI Agent vs. Dashboard vs.
          Personalized Narrative ........................................ 183
    5.4.4 AI Productivity Agent Performance .......................... 186
  5.5 Strengths & Weaknesses ........................................... 190
6 Contextual AI Journaling

6.1 Introduction .................................................. 204
6.2 Related Work .................................................. 209
6.3 Methodology .................................................... 211
  6.3.1 Study Design .............................................. 212
  6.3.2 Demographics ............................................. 214
  6.3.3 Mobile Sensing based Behavioral Data .................... 215
  6.3.4 Personalized Journaling Prompts ........................ 216
  6.3.5 Context-aware Check-ins ............................... 218
6.4 Results ......................................................... 220
  6.4.1 User Study ................................................ 220
  6.4.2 Contextual Journaling Prompts and Check-ins .......... 222
  6.4.3 Journaling Responses Deeper Dive ....................... 230
  6.4.4 Exploring Changes in Wellbeing and Emotional Growth ... 238
  6.4.5 MindScape App Performance ............................ 245
  6.4.6 Strengths, Weaknesses and Improvement Opportunities .. 251
6.5 Discussion .................................................... 257
  6.5.1 Summary of Results .................................... 257
  6.5.2 Implications ............................................. 261
  6.5.3 Limitations and Future Work ............................ 262
6.6 Ethical Considerations ....................................... 264
6.7 Conclusion ..................................................... 265
7 Conclusion

7.1 Future Research Directions ........................................ 268
7.2 Final Comments ....................................................... 269

8 Refereed Publications as a Ph.D. Candidate ...................... 271

A MoodCapture – Surveys and Results .............................. 279
A.1 Surveys .............................................................. 279
A.2 Additional Metrics ................................................. 279

B Productivity Solutions – Surveys and Sample Screens .......... 281
B.1 User survey questions .............................................. 281
B.2 User Experience ..................................................... 283
B.3 Prompts for Extracting Data ....................................... 284
B.4 Prompt for Transforming Data ..................................... 286
B.5 Prompt for Personalized Narrative and Interactive Chat Agent .... 288
B.6 Demographic distribution of user survey. ......................... 290
B.7 Demographic distribution of second phase of the study. ........ 291
B.8 User Survey .......................................................... 292
B.9 Sample Responses from AI Productivity Agent .................. 296
B.10 Sample Personalized Narrative ................................... 297

C MindScape – Surveys and Prompts .................................. 298
C.1 Surveys and Questionnaires ........................................ 298
C.2 Prompt for Check-ins ............................................... 303
C.3 Prompt for Weekend Journal ....................................... 306
C.4 Sample Journaling Prompts ....................................... 308
C.5 Sample Check-ins ................................................... 311
List of Tables

2.1 Participant demographics of College Experience Study . . . . . . . . . 24
2.2 Passive sensing features collected from participants . . . . . . . . . 25
2.3 EMA Self-reports collected from participants . . . . . . . . . . . . . 27
2.4 Comparison between the first and final terms of the participants . . 45
2.5 Comparison between female and male participants . . . . . . . . . 49
2.6 Comparison between on-campus and off-campus students . . . . . 54
2.7 Behavioral changes during COVID-19 and the transition back to in- 

person instruction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58

3.1 Features generated from the mobile sensing system . . . . . . . . . 86
3.2 Changes in behavior and physiology of workers after promotion . . 91
3.3 Differences in patterns of non-promoted/promoted workers . . . . 96
3.4 Changes in patterns after promotion when considering gender . . . 97
3.5 Changes in patterns of higher/lower job performers . . . . . . . . . 100
3.6 Streams that change the most . . . . . . . . . . . . . . . . . . . . . 104

4.1 Demographics, smartphones, and image composition in our study. . 131
4.2 Description of OpenFace features . . . . . . . . . . . . . . . . . . . 137
4.3 Characteristics of the images captured . . . . . . . . . . . . . . . . 142
4.4 Performance of machine learning models on detecting depression . . 143
C.3 Performance .................................................. 299
C.4 Relevancy & Deeper Reflections ..................... 300
C.5 Security & Privacy ........................................... 301
C.6 Satisfaction .................................................... 301
C.7 App and Study Feedback along with open-ended questions. ............ 302
List of Figures

2.1 Data availability for the College Experience Study . . . . . . . . . . . . . 28
2.2 Nullity correlation between the sensing features . . . . . . . . . . . . . 30
2.3 Temporal progression of self-esteem . . . . . . . . . . . . . . . . . . . 32
2.4 Campus events and PHQ4 scores . . . . . . . . . . . . . . . . . . . . 33
2.5 Sleep initiation patterns . . . . . . . . . . . . . . . . . . . . . . . . . 35
2.6 Student anxiety patterns . . . . . . . . . . . . . . . . . . . . . . . . . 36
2.7 Student stress in Summer . . . . . . . . . . . . . . . . . . . . . . . . 38
2.8 Students’ partying/social patterns . . . . . . . . . . . . . . . . . . . . 39
2.9 COVID-19 concern over time . . . . . . . . . . . . . . . . . . . . . . . 42
2.10 Propensity score matching . . . . . . . . . . . . . . . . . . . . . . . . 51

3.1 Demographic information of the promoted participants. . . . . . . 83
3.2 Comparison of changes after promotion . . . . . . . . . . . . . . . . . 94
3.3 Changepoint detection for several sensing streams . . . . . . . . . . . 101
3.4 Composition of different groupings of streams towards changepoint de-
tection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
3.5 Performance of models based on feature selection . . . . . . . . . . . 107
3.6 Performance of machine learning models with and without the use of
ROCKET . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
3.7 Distribution of coefficients of the Logistic Regression model . . . . . . 110
4.1 MoodCapture Framework ........................................ 120
4.2 PHQ-8 application screens ....................................... 130
4.3 PHQ-8 score statistics ............................................. 133
4.4 Objects detected in the background of the images captured .... 141
4.5 SHAP plots describing the top 10 features ...................... 146
4.6 Random forest performance on sub-populations divided by gender and race ............................................. 147
4.7 Comfort Level: Participant’s comfort with the automated capture of their photos........................................ 148

5.1 Study overview ..................................................... 167
5.2 Enhanced Telemetry Data Scraping from Viva Insights using GPT-4 . 169
5.3 Designing productivity solutions .................................. 173
5.4 Participants’ preferences in productivity areas and language ......... 178
5.5 Participants’ preferences in personality traits and social roles ........ 180
5.6 Participants’ overall experience ratings ............................ 184
5.7 Anticipated usage of the distinct data presentation methods ........ 185
5.8 Sample Conversation ............................................. 188

6.1 MindScape Study Design Overview .............................. 212
6.2 MindScape App Workflow ........................................ 216
6.3 Prompt Template for Weekday Journaling ....................... 219
6.4 Prompt Categories and their Distribution ......................... 224
6.5 Check-in Prompt Topics .......................................... 225
6.6 Repetitions of Prompts and Check-ins ........................... 229
6.7 Generic Journal Themes .......................................... 232
6.8 Most Frequently Occurring Emotions ............................ 237
6.9 Improvement in Well-Being as Indicated by EMAs .............. 244
6.10 System Usability Score ........................................ 246
6.11 Participants’ Experience with MindScape ................. 248

B.1 Sample Responses .............................................. 296
B.2 Sample Personalized Narrative ............................... 297

C.1 Example check-ins generated by MindScape. ............. 311
C.2 Example gratitude and self-compassion inducing prompts generated by MindScape. ................................. 312
Chapter 1

Introduction

Section 1.1

Overview

The integration of technology into our daily lives has opened avenues for understanding human behavior in unprecedented detail. With the widespread adoption of smartphones, wearables, and other digital devices, we have access to vast amounts of data that can reveal insights into our behavior, habits, and preferences. This digital footprint provides a unique opportunity to study human behavior in real-time, transcending the conventional data collection methods like surveys and self-reports, which have traditionally been our window into human psyche and behaviors.

Surveys and self-reports, while useful, come with inherent limitations (265). Individuals may not always provide accurate accounts of their behaviors or feelings, either due to memory biases, the desire to present oneself in a favorable light, or simply misunderstanding their own actions and motives. These methods are subjective and vulnerable to various biases including social desirability and recall errors, setting the stage for discrepancies between reported and actual behaviors. To bridge the gap between subjective self-reports and observable realities, researchers have increas-
ingly turned to objective data drawn from the sensors embedded in our everyday digital devices. This shift towards using objective measures—such as sensor data from smartphones and wearables—provides a clearer, more detailed view of human behavior, untainted by the biases inherent in self-reported data. Objective data allows researchers to validate the information gathered through traditional methods and uncover deeper, sometimes hidden, layers of human behavior.

In the domain of human-computer interaction (HCI), a specific focus is placed on behavioral sensing, which leverages the sensor data from digital devices to observe, record, and analyze human behavior in the flow of everyday life. By applying advanced machine learning and artificial intelligence techniques to this data, behavioral sensing can detect patterns and correlations that might be invisible in self-report data alone. This methodological approach has far-reaching implications for fields such as mental health, education, and workplace productivity, where understanding human behavior is critical for intervention and improvement. Consider the ubiquitous smartphone: A repository of data on how we communicate, who we interact with, our physical movements, and even our leisure habits. Wearables monitor our physical activities, sleep patterns, and through biometric sensors, can provide data on our physiological responses that might indicate stress or emotional states. Laptops and desktops, often used for work and education, yield insights into productivity patterns, user attention, and cognitive loads during different tasks. Smartphones, wearables, and laptops are just a few examples of devices that can provide valuable insights into human behavior. Further, one of the most potent aspects of behavioral sensing is its ability to complement subjective methods traditionally used in psychological assessments, such as surveys that capture an individual’s mood or mental state at any given moment. When this subjective data is combined with the objective, concrete data gathered from sensing devices, it gives us a comprehensive view of an individ-
ual’s behavior and mental health over time. This dual-lens approach allows for robust analysis and applications. For instance, predictive models can use behavioral data to anticipate mental health declines or identify when someone might benefit from an intervention, enhancing the capacity for proactive mental health support (233). In workplaces, similar models can predict dips in productivity, offering insights that could be used to adjust workloads or schedules to optimize performance (220).

The incorporation of AI with these sensing technologies is set to revolutionize our approach to both health maintenance and workplace management. AI algorithms, trained on extensive behavioral data, enhance the precision and effectiveness of predictive models and also help us create personalized interventions, turning reactive tools into proactive solutions in managing health and productivity. The research conducted as part of this thesis spans diverse groups—from students and office workers to individuals with clinical conditions like depression—showcasing the wide-ranging applications and transformative potential of behavioral sensing with AI. In the following sections, we will explore recent innovations and advancements in harnessing the power of behavioral sensing and AI to enhance our understanding of human behavior and improve our lives.

Section 1.2 Innovations in Human Behavior Analysis

We present the specific problems and innovations highlighted in this thesis, exploring three major advances in behavioral sensing: 1) Conducting longitudinal sensing studies to gain a long-term understanding of the impact of specific events, such as COVID-19 or job promotions, on the mental health, well-being, and performance of college students and information workers, as explored in Chapters 2 and 3; 2) Utilizing a novel smartphone-camera-based sensing system to reveal clinical insights into...
the mental health and depression states of a clinical population, which is detailed in Chapter 4; and 3) Applying Large Language Models (LLMs) together with behavioral sensing technology to enhance the efficiency and well-being of information workers and students, providing timely interventions in the form of reflective prompts, as discussed in Chapters 5 and 6. Our detailed discussion is as follows.

1.2.1. Capturing Life Events and Transitions with Longitudinal Behavioral Sensing

Navigating through critical life stages such as academia and professional advancements significantly impacts an individual’s mental health and behavior. The intrinsic value of understanding these transitions can not be overstated, particularly when informed by longitudinal behavioral sensing. This approach not only offers granular insights into daily life dynamics but also unveils longitudinal trends that are crucial for designing supportive interventions. We utilize two pivotal studies - each leveraging longitudinal data to illuminate the nuances of life transitions in academic and professional settings.

The *College Experience Study* represents a landmark in longitudinal mobile sensing research.Spanning five years, in this study we track two cohorts of Dartmouth College students, capturing a comprehensive picture of their behavioral and mental health adjustments throughout the undergraduate experience - from pre-pandemic normalcy through the height of COVID-19 disruptions and into the subsequent return to new normals. This unparalleled duration allows for acute observation of resilience patterns and mental health trajectories, underscoring the extended impact of external stressors such as a global pandemic on student life. Crucially, the data reveals persistent shifts in behavior alongside reversion to pre-pandemic norms, highlighting the dual nature of human adaptability in crisis situations. The study’s longitudinal breadth provides a rich dataset that contributes significantly to our understanding of mental health and
resilience among college students, emphasizing the importance of supportive academic environments.

Complementarily, study on detecting job promotion explores the professional domain, analyzing the aftermath of job promotions over a year-long period. This study highlights the inherent complexities of professional advancements, capturing a detailed account of physiological and behavioral changes post-promotion. By utilizing passive mobile sensing data, the research offers objective insights into the adaptive processes demanded by new professional responsibilities and roles, marking a significant departure from conventional self-report methods. The findings unveil gender-specific stress patterns and illustrate the broader implications of career progression on personal well-being. This longitudinal analysis shed light on the intricate balance between career growth and mental health, paving the way for discussions on workplace wellness.

Both studies showcase the strength of using longitudinal mobile sensing to capture how people navigate through major life changes. They demonstrate that by looking at behavior over time, we can uncover patterns that short-term studies might miss. This is key for creating programs and supports that truly meet people where they are, helping them adapt over time to shifts in their personal and professional lives. In addition, the combination of objective data with personal stories in these projects presents a powerful method for deeper understanding. It is this blend of numbers and narratives that can fuel better support systems for everyone from students facing the pressures of college to employees stepping into new roles.

In essence, the cumulative findings of these studies advance academic discourse and offer practical guidelines for enhancing well-being across different life stages. They underscore the unique advantages of employing longitudinal methodologies in behavioral sensing research, setting a new standard for future investigations into the
human experience.

1.2.2. Assessing Mental Health via In-the-wild Smartphone Images

In an era where mental health disorders, particularly depression, present a growing public health concern, the quest for innovative detection and monitoring solutions is more crucial than ever. Given that over 264 million people worldwide grapple with depression, leading to significant impacts on overall health, social relationships, and work productivity, there is an urgent need for early detection and intervention (249).

The MoodCapture Study introduces a pioneering approach by analyzing naturalistic images captured during daily phone use. This method stands out because, unlike traditional surveys or clinical evaluations, it captures genuine expressions in real-time. This is a big leap forward, offering a more consistent and unobtrusive peek into how people are feeling.

We collect over 125,000 photos from 177 participants, all diagnosed with major depressive disorder, over 90 days. These are not just any photos—these are snapshots of real moments, taken as participants go about their days. What we find is compelling: certain image attributes, like the angle of the phone or the lighting in the room, could help identify depressive states. By training a machine learning model with these image characteristics, we could effectively classify if someone was experiencing symptoms of depression. By focusing on in-the-wild images, the project also touches on a crucial aspect of tech-driven health care: privacy. Discussing participants’ comfort levels and privacy builds trust in this new approach and emphasizes the need for tech solutions that respect personal boundaries.

MoodCapture exemplifies the advancements we are making in using technology to understand and support mental health. It showcases the power of AI and behavioral sensing in creating effective, user-friendly tools for early depression detection and ongoing mental health monitoring, promising a future where help is available at the
1.2 Innovations in Human Behavior Analysis

Introduction

touch of a button, tailored to each individual’s life and needs.

1.2.3. Enhancing Personal Productivity and Well-being through LLM and Behavioral Sensing Integration

The fusion of LLMs with behavioral sensing presents a new frontier for enhancing personal productivity and well-being. Our investigation explores personalized productivity enhancement and contextual AI-powered journaling to demonstrate the vast potential of these integrations.

The "From User Surveys to Telemetry-Driven Agents" project highlights our journey from understanding user preferences through surveys to implementing these insights in developing a GPT-4 (246) powered productivity agent. This agent, informed by real-world telemetry data, showcases the effectiveness of LLMs in creating solutions that truly resonate with the user’s day-to-day activities. By comparing this AI-driven approach with traditional productivity tools, we uncover the significance of adaptability, personalization, and privacy in designing future AI-assisted productivity tools.

We first try to understand the intricacies of user preferences in productivity tools through an extensive survey involving 363 participants. With the insights from the first phase, we develop a personalized productivity agent powered by GPT-4, a state-of-the-art LLM known for its contextual understanding and generative capabilities. It utilizes telemetry data from Viva Insights, encompassing email and meeting patterns of information workers, to tailor its assistance. This customization is rooted in the foundational belief that productivity tools must adapt to individual workflows to be genuinely effective. We then perform a comparative study with 40 participants, juxtaposing the AI agent with other productivity tools like dashboards and narratives, underscoring the importance of user-centric design.

Further extending the exploration, "MindScape" presents a novel integration of
behavioral data with LLMs to promote mental health through journaling. Targeting college students, a demographic often burdened with unique pressures, we use mobile sensing to inform the generation of personalized, context-aware journaling prompts. This approach explores the power of LLMs to analyze and interpret vast amounts of personalized data, facilitating a tool that offers meaningful mental health support tailored to the individual’s daily experiences and emotional states. Building on preliminary insights from a user study, "MindScape" demonstrates the effectiveness of embedding behavioral intelligence within AI applications. The resultant application offers nuanced mental health support, blending seamlessly into users’ lives, encouraging reflective practices critical for emotional development and resilience.

Both projects underscore the significant advancements in utilizing AI and LLMs when combined with behavioral sensing. They offer powerful examples of how these technologies can create deeply personalized experiences that resonate with users’ unique needs and situations. While "From User Surveys to Telemetry-Driven Agents" focuses on optimizing work efficiency through tailored assistance, "MindScape" focuses on fostering mental well-being through reflective journaling, guided by processed behavioral cues.

This synergy between AI and behavioral data is not just innovative; it represents a paradigm shift in designing digital tools for personal development. By capturing and analyzing real-time data, whether for improving productivity or supporting mental health, we harness the potential to offer solutions that are not just reactive but proactive, adapting to changes in user behavior and emotional states over time. The insights gained from these initiatives highlight a future where digital solutions are intimately aligned with our preferences and emotional landscapes, showcasing the tremendous promise of behavioral sensing and AI in advancing personal development tools to new heights of intuitiveness and responsiveness. This journey towards integrating AI
with behavioral sensing sets a benchmark for future explorations, emphasizing AI’s capacity to significantly improve the quality of human life and productivity.

**1.3 Protection of Human Subjects**

All the studies presented in this thesis have received approval from the Institutional Review Board (IRB) of the respective institutions. The Tesserae Study obtained approval from the University of Notre Dame’s IRB and via reliance agreements with other participating research universities. The College Experience, MoodCapture, and Contextual AI Journaling studies received approval from the Committees for the Protection of Human Subjects at Dartmouth College. Likewise, the study focusing on personalized productivity solutions was approved by the IRB of Microsoft Research.

We have implemented several strategies to ensure the security, privacy, and well-being of our participants. All collected data, including sensor readings, survey responses, and Ecological Momentary Assessment (EMA) data from smartphones, are securely stored on servers. Data from smartphones are transferred to these servers periodically using the Transport Layer Security (TLS) standard to ensure their safe transmission. To access the servers, one needs a password-protected login, and these servers are physically located in offices that can be locked. Instead of using names, the data on our servers are assigned research-specific identifiers. Furthermore, any documents that contain personally identifiable information, such as signed consent forms and contact information sheets necessary for project management, are kept in a separate location, accessible only to authorized members of the project team. Periodic assessment interviews and measures conducted by our study staff are also documented using identifiers instead of personal names. Access to this data by researchers not part of our team is restricted according to IRB protocols and participant
We have obtained informed consent from all participants, clearly explaining the study’s purpose and how their data will be used. We have emphasized their right to withdraw from the study at any moment without any repercussions. To ensure confidentiality, all data used for analysis were anonymized and de-identified. This involved stripping away all personally identifiable information through a rigorous anonymization process. Specifically, for images captured in the wild, we implemented a protocol where any photograph inadvertently capturing individuals or nudity was flagged during a review by two team members and was promptly removed. Access to the data is stringently restricted to authorized personnel only. Where applicable, images were processed to anonymize identifiable features without compromising the integrity of the data necessary for our research. We recognize the importance of addressing concerns related to the use of digital data, particularly regarding privacy and security. To this end, we have diligently followed applicable data protection regulations and adhered to best practices to safeguard the information we collect and manage.

The Tesserae Study discussed in this dissertation received partial support from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under IARPA Contract No. 2017-17042800007. The College Experience Study was funded by the National Institute of Mental Health (NIMH) via grant number 5R01MH059282. Additionally, the MoodCapture study received support from the NIMH, under award number R01MH123482-01. The research on personalized productivity solutions and contextual AI journaling was funded internally.
In this thesis, we make several distinct contributions to the fields of behavioral psychology, mental health, human-computer interaction and artificial intelligence, broadening the understanding and application of technological tools in both personal and professional domains. Below are the key contributions of this research:

(a) We present the College Experience Study, which is the longest continuous mobile sensing study to date. This study contributes a detailed longitudinal analysis of college students over four pivotal years, providing deep insights into how events like the COVID-19 pandemic impact student mental health and behavior. By utilizing mobile sensing technology and Ecological Momentary Assessments (EMA), we uniquely capture both the quantitative and qualitative aspects of student life, revealing critical factors such as increased stress during transitional periods back to campus and significant behavioral changes influenced by the pandemic. We also explore gender-specific behavioral trends and the effects of changing living arrangements. Moreover, we release a comprehensive dataset to enrich the field, fostering further research on the long-term mental health trends of college students and offering substantial evidence for developing targeted interventions to support student well-being during crises. These contributions underscore the transformative potential of longitudinal mobile sensing in understanding and addressing the complexities of mental health in academic environments.

(b) We deliver pioneering insights into how job promotions affect physiological and behavioral changes among information workers, significantly enhancing our understanding of workplace dynamics and individual adjustment. For the first
time, this study deploys mobile sensing technology to capture real-time, objective data including sleep patterns and physical activity before and after job promotions. Key findings highlight gender-specific reactions to promotions; for instance, promoted female workers experience extended stress durations during work hours unlike their male counterparts. Additionally, employing advanced data analytics for feature extraction from multivariate time series demonstrates the capability of mobile sensing streams to detect job transitions, achieving notable success in classification performance. This research underscores the substantial influence of job promotions on individual workers, and lays the groundwork for developing supportive measures to help workers adapt to new professional roles, thereby promoting enhanced well-being and productivity in the workplace.

(c) We build and deploy the MoodCapture system, an innovative mobile app that passively captures unobtrusive, in-the-wild smartphone images to detect depressive symptoms. The application actively captures authentic emotional states through spontaneous front-facing camera images, enhancing the ecological validity of mental health assessments beyond traditional methods. Our key contributions include analyzing image characteristics such as illumination, location, and phone angle, which are crucial for AI-driven health diagnostics. The study employs machine learning models, notably a random forest algorithm using 3D face landmarks, achieving an improved accuracy in depression detection over baseline measures. Additionally, the research gathers user acceptance data, providing vital insights into privacy concerns and user comfort, which will guide the design of future non-invasive mental health tools. By integrating nuanced user interaction data with psychological health assessments, this pioneering research advances mobile health diagnostics and makes significant contributions.
to the fields of Human-Computer Interaction and affective computing.

(d) We explore the development and impact of a personalized productivity agent leveraging state-of-the-art GPT-4 LLM, informed by extensive user surveys and telemetry data analysis. The study begins with a detailed survey of 363 information workers to uncover specific user needs and preferences for productivity tools, revealing a strong preference for proactive suggestions and comfort with sharing high-level telemetry data for personalized responses. Utilizing these insights, we develop a unique application that facilitates the sharing of user telemetry data, enabling the AI-based agent to provide customized assistance. This agent, powered by GPT-4, dynamically tailors its engagements by extracting pertinent information from workplace telemetry data, enhancing user productivity. We conduct a rigorous evaluation involving 40 participants to assess the agent’s effectiveness compared to conventional data presentation methods like dashboards and narrative summaries. Through this comprehensive study, we demonstrate the transformative potential of AI in creating personalized, context-aware productivity tools that respond adaptively to individual user needs, thereby setting a new standard in AI-driven workplace efficiency enhancement.

(e) We build and evaluate MindScape system, a pioneering tool combining mobile sensing and LLM to produce personalized journaling prompts for college students. By incorporating real-time behavioral data, MindScape tailors prompts that significantly enhance user engagement and mental well-being. Over eight weeks, 20 student participants report an 11% improvement in well-being scores, with notable enhancements in areas such as affect, loneliness, mindfulness, self-reflection, anxiety, and depression. The study analyzes the prompts for topic relevance and linguistic style, establishing that responses to contextual prompts
involved more personal language and deeper reflection. Additionally, user feedback highlight the app’s usability and the relevance of the prompts, with 85% of participants praising its usability and 70% affirming the prompt’s relevance. This comprehensive study demonstrates the potential of integrating LLMs and behavioral sensing to enrich mental health support through personalized digital journaling.

The thesis is structured as follows: The first part (Chapters 2-3) focuses on advancements in behavioral sensing through longitudinal studies. Chapter 2 presents a four-year study, the longest mobile sensing study to date, involving college students. This chapter explores the modeling of student behaviors across this period. Chapter 3 details a year-long study among information workers, examining behavioral differences before and after promotions, and investigating correlations between smartphone data and mental health. In the second part of the thesis (Chapters 4-6), we apply behavioral sensing methods in innovative ways. Chapter 4 introduces the novel application of passively captured smartphone images to detect depression in participants. Chapters 5 and 6 leverage recent advancements in AI, specifically in LLMs. Chapter 5 focuses on creating tools that integrate behavioral sensing to enhance workers’ productivity and efficiency. In contrast, Chapter 6 aims to improve the mental well-being and personal growth of college students with the help of behavioral sensing and AI. Overall, the studies, methods, and results presented in this thesis pave the way for advancing behavioral sensing with AI, potentially leading to more efficient and healthier lifestyles in the future.
Chapter 2

Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic

Section 2.1

Introduction

The college years represent a critical period in students’ lives, marked by significant personal, social, and academic growth (195; 338; 12). Understanding the behavioral patterns and mental health of college students during this time is essential for promoting their well-being and success. Data-driven approaches, such as mobile sensing techniques and Ecological Momentary Assessment (EMA) surveys, have
emerged as valuable tools for investigating the link between daily activities and mental health (130; 84; 325). By analyzing sensor data and survey responses, researchers can uncover trends in students’ daily activities, behavioral trends, social interactions, and emotional states, shedding light on the factors that influence their mental health and overall well-being. The landscape of college life has dramatically shifted in recent years due to the COVID-19 pandemic and subsequent national lockdowns (298; 283). These changes have presented unprecedented challenges for students, highlighting the need to better understand how they affect mental health. Longitudinal studies that examine students’ behaviors and experiences throughout their college career, encompassing both the pre-pandemic years and the impact of COVID-19, can provide a deep understanding of the factors that contribute to their mental health and resilience. Such studies would offer valuable insights for supporting students during both normal and extraordinary times.

In this chapter, we present the findings of a comprehensive, longitudinal study that leverages both passive sensing data and EMA surveys to explore the experiences of two cohorts of undergraduate students at Dartmouth College over a period of five years, spanning from September 2017 to June 2022. Our analysis encompasses various aspects of college life, including pre-pandemic years, the COVID-19 pandemic, and the gradual return to normalcy as the pandemic subsided. By examining behavioral patterns and mental health metrics during different periods, we aim to uncover the factors that contribute to students’ well-being and resilience, and to provide valuable insights that can inform the development of effective interventions and support systems for college students. This study utilizes the StudentLife mobile application (329) used and validated in many previous studies (34; 234; 138; 333; 230) to facilitate the consistent and fine-grained collection of passive sensing data and EMA surveys from undergraduate students over an extended period. By gathering objective data at
2.1 Introduction

Capturing the College Experience

varying intervals, in addition to weekly self-reported EMA surveys, we are able to paint a detailed picture of students’ daily activities, behaviors, mental health, and overall college experience. The rich dataset provided by the sensing app allows us to explore various aspects of college life, such as behavioral differences based on gender, changes in living arrangements (on-campus vs. off-campus), and the persistence of behavioral patterns after the pandemic subsides.

Our study not only examines the impact of the COVID-19 pandemic on students’ mental health and daily activities but also delves into the broader college experience over four years. By analyzing pre-pandemic data, we can identify trends in students’ mental health and behavior prior to COVID and across their college years. This comprehensive approach enables us to understand the resilience and adaptability of students in the face of unprecedented challenges, as well as the factors that may contribute to their mental well-being. In particular, we found evidence that events during the COVID-19 pandemic influenced students’ concern levels over time, as demonstrated by the link between data evaluating concern and mental status and events. We also observed differences in behavioral patterns between the first and last academic terms of both cohorts, highlighting changes in behavior over the course of students’ college years. The first cohort of $N = 104$ students started college in Fall 2017 and graduated in Spring 2021, while the second cohort of $N = 113$ students started in Fall 2018 and graduated in Spring 2022. We also identified differences in behavioral patterns in different phases (pre-COVID, during-COVID and in-person instruction which is the period when the students returned to campus after COVID subsided), suggesting the effect of the pandemic on human behavior change. In addition to extensive quantitative analysis of data collected through our mobile sensing app, we also conducted qualitative interviews with fifteen students who graduated in 2022 and 2023. These interviews provide additional insights into the impact of the
COVID-19 pandemic and other factors on college students’ mental health and daily activities. By integrating both quantitative and qualitative data, our study offers a comprehensive understanding of the interplay between daily activities, mental health, and resilience during the college years. This mixed-methods approach allows us to identify patterns and trends in the data while capturing the richness and diversity of students’ experiences. The chapter makes a number of contributions including:

(a) **Longitudinal Study and Analysis of the College Years**: Our study presents a rare longitudinal analysis of undergraduate students, offering valuable insights into their mental health and daily activities over an extended period. We visualize the sensing data and self-reports that show insightful trends, such as a spike in student stress and anxiety levels in the run-up to returning to campus from breaks. This increase might be tied to the impending return to campus and academic burdens. Our study and analysis shed further light on the challenges students faced during their college years, including the rising risk during the COVID-19 pandemic.

(b) **Factors Influencing Student Mental Health**: We delve into the influence of various factors on students’ mental health and their overall college experience. We examine behavioral differences based on gender, comparing the first term to the final term, and consider changes in living arrangements. For example, we find that female students exhibit higher sedentary behavior, elevated COVID concern, higher social level, but lower self-esteem. We also identify significant behavioral and mental shifts between the first and last academic terms, highlighting differences in behavior between the initial and final semesters.

(c) **Impact of the COVID-19 Pandemic on College Life**: We explore the persistence of behavioral patterns after the COVID-19 pandemic subsides, aiming to provide insights for the development of targeted interventions and support
systems. Our study reveals a noticeable negative impact of the pandemic on students’ mental health, with increased levels of depression, anxiety, stress, and psychological distress. While certain behaviors, such as physical activity and time spent at specific locations, are returning to pre-pandemic levels, others, such as sleep duration, remain elevated, and social interaction is reduced. These findings emphasize the need for additional support during the transition back to normal routines.

(d) Qualitative Insights from Undergraduates: To deepen our understanding of the college experience during this unique period, we incorporate qualitative interviews with graduating students. These interviews offer insight into the challenges and successes students encountered while navigating their college years amid a global pandemic.

(e) College Experience Study Dataset and Open Research Questions: In the interest of advancing research into the mental health of college students during their college years using mobile sensing and self-reports, we discuss the public release of the de-identified College Experience Study dataset used in this chapter. We also outline a series of open research questions that could be studied using the public dataset. The College Experience Study dataset is available here (142).

The College Experience Study is the longest mobile sensing study to date, investigating the relationship between daily activities, behaviors, mental health, and resilience during the college years. The findings have significant implications for interventions, policies, and support systems promoting student well-being and success in higher education, especially during a global pandemic. In this study, our primary focus is on examining the experiences of college students at a single institution in the US, recognizing that our findings may be uniquely applicable to this specific demographic
group. As a result, we underscore the need for further research to assess the relevance of our insights across various populations and settings.

This chapter presents related work, background information, and an overview of the dataset in Sections 2.2 and 2.3. It explores the collected data, including the impact of COVID-19, in Section 2.4. Detailed observations on differences between the first and final term, gender, changes in living arrangements, and persistent behavioral patterns during the pandemic are provided in Sections 2.5, 2.6, 2.7, and 2.8, respectively. We discuss the findings, implications and ethical considerations in Section 2.9. This is followed by limitations and future work in Section 2.10. The public release of the de-identified dataset and open research questions that could be studied with the dataset are discussed in Section 2.11. We conclude in Section 2.12 with reflections on the study and some final remarks.

Section 2.2

Related Work

Mobile sensing has surfaced as a prevalent tool for passively analyzing human behavior (1; 176; 259; 185). A considerable volume of research has established connections between mobile sensing and diverse facets of personality (339; 277), mental health (280; 225; 3; 2; 27; 92; 255), and professional behavior (220; 79; 232; 230; 207; 231; 172; 254). This encompasses associating daily activity with academic performance (329; 280; 330), identifying mental health risks (4; 334; 95), and modeling behavior for mental health assessment (184; 194; 307; 213). Recently, data-driven methodologies for examining the relationship between daily activities and mental health have attracted significant scholarly interest (43; 222; 66). Mobile sensing techniques, together with Ecological Momentary Assessment (EMA) surveys, have been widely employed to gather behavioral data, thereby formulating models
that encapsulate mental health parameters, specifically targeting student demographics (349; 335; 194; 211; 212; 287). The COVID-19 pandemic has underscored the potential of mobile sensing and data analytics for tracking the progression of diseases at various scales—individual, neighborhood, city, and national levels. Mobile sensing has evolved into an indispensable resource for identifying personal infection statuses, monitoring long-term health shifts, tracing community epidemics, and observing the development of viruses and their variants (336). The passive, in-situ data gathering facilitated by mobile sensing offers the distinct advantage of enabling factual deductions in real-world settings, thus proving to be a highly effective tool for studying human behavior during the pandemic. The insights derived from this methodology considerably augment our comprehension of how individuals navigate such extraordinary situations.

The COVID-19 pandemic and ensuing lockdowns have significantly affected students’ lifestyles and mental well-being (279; 58; 358). While some students have managed to quickly adapt, others have grappled with stress, anxiety, and other mental health issues during this period (348; 234; 138). Prior research has explored the correlation between daily activities and mental health by employing statistical and machine learning techniques to scrutinize time-series data during (305; 217) and before the pandemic (329; 194; 56). However, very few of these investigations can be termed as long-term tracking studies. Our research leverages a mobile sensing app that facilitates persistent, granular data collection over years, permitting remote and cost-effective capture of continuous sensor data from devices and self-reported EMA surveys. This chapter presents the findings of a substantial longitudinal multi-year study spanning nearly five years involving two cohorts of 104 and 113 undergraduate students from Dartmouth College, a four-year Ivy-league University in the United States. Our study gathered up to 20 different sensory features at varying intervals.
and responses to numerous surveys weekly, offering a comprehensive understanding of college life and the COVID-19 pandemic’s impact on students’ daily routines and mental health.

Numerous existing studies provide insights into behavioral changes during the pandemic (336; 214; 234), but few of them concentrate on the return to baseline after COVID subsides. The majority of the studies that explore the behavior and mental health after COVID subsides usually just focus on the pre-post lockdown phases or a specific outbreak period (346; 204; 76). However, our data spans pre-pandemic, pandemic, and the return to in-person instruction, hence, we believe we are ideally positioned to undertake return to baseline analysis. Our research investigates gender-based differences in students and the behavioral discrepancies between their first and last terms at university. We also explore the implications of changing living arrangements during COVID. While several studies examine changes in living arrangements concerning their influence in the spread of the COVID-19 virus (189; 15), they do not focus on the differences between on-campus and off-campus students. It should also be noted that there are more studies that investigate the impact of changing living arrangement and remote learning on students’ behavior and mental health, although they primarily use self-reports only (5; 224). Our research offers invaluable insights into the relationship between daily activities and mental health during the COVID-19 pandemic, informing the development of effective mental health support interventions for students. While similar to our longitudinal data, Xu et al. (352; 350) conducted a four-year study of 497 students using mobile sensing and surveys; each student participated in the study for a 10-week period each year. In contrast, our study maintains a consistent participant group throughout the four years. While we experienced some participant dropout—an expected occurrence in long-term longitudinal studies—we examined the behavioral and mental health changes of the same
students from their college entrance to graduation. This continuity offers a more comprehensive understanding of how students’ behavior and mental health evolve during their college experience, particularly in the context of a global pandemic.

### Section 2.3

**College Experience Study**

In this section, we discuss the details of our study methodology, including its design, collected passive signals and Ecological Momentary Assessments (EMAs).

#### 2.3.1. Study Design

In our study, we employed a mixed-methods approach, specifically utilizing a convergent design with a supplementary qualitative component. This design enabled us to analyze the behavioral sensing data and self-reports quantitatively, while concurrently conducting interviews with fifteen undergraduate students to provide context and clarify the quantitative findings. By combining the behavioral data and EMAs with qualitative insights from students’ experiences, we gained a more comprehensive understanding of the results. Our methodology comprised two phases. In the first phase, we collected mobile sensing data, EMA responses, and a variety of other validated psychometric measures from students over a four-year period. In the second phase, we sought to understand the perceptions and experiences of fifteen students by conducting in-depth interviews. It is important to note that these students, who graduated in 2022 and 2023, were not part of the main study sample. They were selected for their ability to provide valuable contextual information. This streamlined process allowed for efficient data collection and facilitated an interpretive analysis of diverse findings. Incorporating interpretation was crucial in our study, as it offered insightful context and explanations to better understand the relationships between
the quantitative results and the students’ experiences.

Table 2.1: Participant Demographics: The table presented provides a detailed breakdown of the demographic composition of the students involved in our study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>129</td>
<td>69.35%</td>
</tr>
<tr>
<td>Male</td>
<td>57</td>
<td>30.65%</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>110</td>
<td>59.14%</td>
</tr>
<tr>
<td>Asian</td>
<td>43</td>
<td>23.12%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>6</td>
<td>3.23%</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>6</td>
<td>3.23%</td>
</tr>
<tr>
<td>More than one race</td>
<td>15</td>
<td>8.05%</td>
</tr>
<tr>
<td>Not reported</td>
<td>6</td>
<td>3.23%</td>
</tr>
</tbody>
</table>

2.3.2. Recruitment

In this study, a sample of 217 first-year undergraduate students from Dartmouth College in the United States were recruited. The university registrar sent all incoming admitted students information about the four-year study. The recruitment process took place from August 2017 (cohort 2021) to November 2018 (cohort 2022), with students enrolling and providing consent at the beginning of their first academic year. However, two participants were excluded from the dataset and subsequent analyses due to one participant’s mobile phone incompatibility with the sensing application and another withdrawing within the first week of the study. Consequently, the final sample consisted of 215 participants, of which 146 (67.8%) were female. The age range for all participants at the time of enrollment was 18 to 22 years. The study was approved by Dartmouth College’s Institutional Review Board (IRB). Potential participants were initially asked to complete an online survey to determine their eligibility for the study. Upon meeting the eligibility requirements (such as, possession of an Android or iOS smartphone compatible with the study’s mobile application) and expressing an
interest in participation, individuals proceeded to complete a series of online surveys. Subsequently, students were asked to install our data collection application on their Android or Apple smartphone. The mobile sensing application collected two types of data: sensing data to capture users’ behavior and EMA data to measure mental health. Participants received compensation for their weekly EMA responses at a rate of 10 dollars per week. Note that they did not receive any additional compensation for the passive sensing data collection.

2.3.3. Demographics

Out of the 215 participants, 186 students have at least three years of data. Table 2.1 shows the demographics of the 186 students used in our analysis. The majority (69.35%, N=129) of our participants identify as females. In terms of race, 59.14% (N=110) are White, 23.12% (N=43) are Asians, 3.23% (N=6) are Black or African American, 3.23% (N=6) are American Indian/Alaska Native and 8.05% (N=15) belong to more than one race.

Table 2.2: Features: The table below lists the passive sensing data we collect from the participants.

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Activity</td>
<td>Walking / sedentary (still) / biking / running duration</td>
</tr>
<tr>
<td></td>
<td>duration in vehicle, number of steps</td>
</tr>
<tr>
<td>Mobility &amp; Semantic Locations</td>
<td>Distance travelled, time spent at home, workout places,</td>
</tr>
<tr>
<td></td>
<td>study places, social places, dorms, others’ dorm, greek houses,</td>
</tr>
<tr>
<td></td>
<td>max distance from campus, number of locations visited</td>
</tr>
<tr>
<td>Phone usage</td>
<td>Number of phone locks &amp; unlocks, duration of phone unlock</td>
</tr>
<tr>
<td>Audio Plays</td>
<td>Number of audio plays, duration of audio plays</td>
</tr>
<tr>
<td>Sleep</td>
<td>Sleep duration, sleep start time, sleep end time</td>
</tr>
</tbody>
</table>

2.3.4. Sensor Data

Our mobile application taps into various sensors embedded within smartphones to record data, and we further compute derivative features from this raw data. Below, we offer an overview of the types of features we gather. For a complete list, please
Physical Activity. Our application utilizes the Activity Recognition API available on both iOS and Android devices to record the physical activities of the participants. **Mobility and Semantic Locations.** By employing GPS data, we gather information on students’ mobility, which allows us to compute features such as the total distance travelled and the number of unique places visited. Furthermore, we assign semantic labels to various locations on campus (e.g., gym, study place, dorm, social places, cafeteria, classrooms, etc.). Since most buildings on the university campus serve a single purpose, this enables us to analyze student behavior in different contexts. **Phone Usage.** As an indicator of screen-time and overall phone usage, our application tracks the number of times students lock and unlock their phones. We compute the total number of phone lock and unlock events and the average duration between these events. **Audio Plays.** Our sensing application monitors whether any audio session, encompassing any audio-based media like music or video, is active on the phone. We compute the number of audio plays and the total duration of audio play. **Sleep.** We estimate bedtime, wake-up time, and sleep duration using the method described in (63; 329). Please note that the sleep data has a measurement error of approximately +/- 32 minutes.

For analytical and interpretive ease, we categorize each of these feature groups into four periods: epoch 0 (overall, encompassing a full 24-hour period), epoch 1 (morning, 12 am to 9 am), epoch 2 (day, 9 am to 6 pm), and epoch 3 (night, 6 pm to 12 am).
2.3.5. Ecological Momentary Assessments

We collect weekly EMA self-reported surveys from the participating students. The categories of responses are shown in Table 2.3. We use the EMA responses to define the mental health outcomes in subsequent analysis. Note that the first two questions of Patient Health Questionnaire-4 (PHQ4) are used as anxiety questions, and the last two capture depression.

2.3.6. Data Completeness, Compliance and Challenges

We assess the consistency and quality of our data in this section, focusing on two main components: sensor data collected from smartphones and responses to EMAs. For the Table 2.3: Self-Report Measures: The following table enumerates the EMA self-reports collected from participants, delivered randomly once per week through our mobile application.

<table>
<thead>
<tr>
<th>Self-reports</th>
<th>Median</th>
<th>Std.</th>
<th>Range</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COVID</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How concerned are you about COVID-19?</td>
<td>4.00</td>
<td>0.49</td>
<td>1-7</td>
<td></td>
</tr>
<tr>
<td><strong>Patient Health Questionnaire-4 (PHQ4)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over the last 2 weeks, how often have you been bothered by the following problems?</td>
<td>2.00</td>
<td>2.06</td>
<td>0-12</td>
<td></td>
</tr>
<tr>
<td>Feeling nervous, anxious or on edge;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not being able to stop or control worrying;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feeling down, depressed or hopeless;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little interest or pleasure in doing things;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Self Esteem</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right now, I worry about what other people think of me.</td>
<td>2.00</td>
<td>0.81</td>
<td>1-5</td>
<td></td>
</tr>
<tr>
<td>Right now, I am pleased with my appearance.</td>
<td>3.00</td>
<td>0.77</td>
<td>1-5</td>
<td></td>
</tr>
<tr>
<td>Right now, I feel as smart as others.</td>
<td>3.00</td>
<td>0.77</td>
<td>1-5</td>
<td></td>
</tr>
<tr>
<td><strong>Social Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have you spent most of your time alone or with others today?</td>
<td>3.00</td>
<td>0.69</td>
<td>1-5</td>
<td></td>
</tr>
<tr>
<td><strong>Stress</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you feeling stressed now?</td>
<td>2.00</td>
<td>0.66</td>
<td>1-5</td>
<td></td>
</tr>
</tbody>
</table>
sensor data, we calculate a completeness ratio, representing the percentage of days with recorded data relative to each participant’s total study duration. The EMA data is evaluated on a weekly basis, comparing the number of weeks a student responded to the EMA against their overall study participation duration. Additionally, we examine the missing rates for each feature to confirm there are no substantial data gaps.

Figure 2.1: Data Availability: Figures (a) and (b) display sensing data and EMA self-reports availability, respectively. The x-axis represents days of sensing data in (a) and weeks of EMA responses in (b); the y-axis indicates the number of students in both. The smooth line in each plot shows the kernel density estimate, with a trend towards data availability for a higher number of days and weeks, approximating the 4-year study duration.

Figure 2.1a illustrates the distribution of the average number of days for which we have available sensing data per participant. On average, we have sensing data available for 1104 days per participant. This translates to an average completeness rate of 90%, implying that for each participant, we have data for 90% of the days from their enrollment date to their exit from the study. Similarly, Figure 2.1b showcases the distribution of the average number of weeks for which we have EMA responses. On average, participants have responded to the EMA for 130 weeks, yielding an average EMA completeness score of 75%. Both plots reveal that the data availability is skewed towards a higher number of days and weeks. The smoothed line on the plot, representing the kernel density estimate, peaks at one point, indicating that we have unimodal data with one common value for days and weeks. In terms of missing data,
we find an average per-user missing value rate of approximately 20% for location-based features, 12% for audio play-based features, and 22% for steps. We have no missing data for physical activity features, and as the phone usage and sleep features are derived, they invariably have non-empty values.

We conducted Little’s Missing Completely At Random (MCAR) test to examine the missing data mechanism in our dataset and obtained statistically insignificant result (p-value > 0.10), suggesting that most variables are likely missing completely at random (MCAR). Little’s MCAR test is a statistical method to determine whether missing data in a dataset is completely random. We also explored the possibility of data being missing at random (MAR) for specific variables by calculating correlations between the binary indicator of missing values for each feature and the observed values of other features in the sensing data. MAR refers to a type of missing data mechanism where the missingness is related to observed variables but not to missing values themselves. Our findings reveal that several location-based features have strong correlations (ρ-value ≥ ±0.40) with missingness indicators, suggesting that location-based features could be MAR rather than MCAR. We further examined missing values using a nullity correlation matrix. A nullity correlation matrix summarizes the relationships between missing data patterns for all pairs of variables in a dataset, where nullity correlation quantifies the strength and direction of the association between missingness patterns for two variables. Figure 2.2 visualizes the relationships between missing values across the behavioral sensing data, showing a strong association (ρ-value ≥ 0.7) across most location-based features. This result aligns with the notion that location-based features might be MAR, implying a systematic pattern of missingness across these features. Various factors could contribute to this finding, such as participants turning off their GPS or location services on their phones due to privacy concerns or battery usage. Alternatively, missingness in
location-based features might relate to the quality of GPS signals or the frequency of data collection, influenced by participants’ behaviors or characteristics. While we lack complete demographic information and other characteristics of the participants, we further analyzed whether students’ demographics play a role in the missingness of each feature. We calculated the missing rate per feature for each participant and used the Kruskal-Wallis test (a non-parametric test to compare independent groups) to compare the missing rates for different demographic groups (race, gender) and cohorts (2021 and 2022) for each feature. Our analysis revealed no significant difference in missingness for features across race and cohorts. However, we found a statistically significant difference (p-value < 0.10) in missingness for the audio plays feature between genders – males had a higher missing rate (18%) compared to females (8%).

Figure 2.2: Nullity Correlation: The figure illustrates the strength of correlation between the presence or absence of one feature and that of another. Most location features have a value of 1, indicating that if one location feature appears, the other almost certainly does as well. Given that location features depend on the availability of GPS data, it’s logical that they would co-occur, either both being present or both absent.
2.3 College Experience Study — Capturing the College Experience

In addition to the missing data patterns observed in our dataset, it is important to consider the challenges and context of conducting a four-year longitudinal study using Android and Apple phones, as these factors have influenced data collection and participant retention. Conducting such a study presents challenges due to the continuous evolution of students’ phones and operating system updates. For instance, between 2017 and 2022, there were six and seven new versions of Apple iOS and Android OS, respectively. Pushing these updates often relied on cellular providers and required modifying and updating our app to ensure proper functionality. In some cases, these updates caused conflicts between the hardware or software, leading to students dropping out of the study. Moreover, even when updates went smoothly, data loss occurred as students did not immediately reinstall the app update after receiving the new OS. Consequently, compliance rates for completing the four-year study were affected by retention issues and technological updates. This situation created an ongoing race for app developers to stay ahead of significant OS changes before they were pushed by providers. One notable change was introduced in Android 10 (released on September 3, 2019), which imposed restrictions on when apps could initiate activities while running in the background. This change had a significant impact on our passive sensing application. Typically, passive sensing apps push APIs, phones, and platforms to their limits, balancing low-power sensing, machine learning inference, and cloud integration while considering energy consumption. Taking these factors into account, out of the initial study size of \( N=215 \) participants, a total of \( N=123 \) students completed the four-year study. Specifically, we have the sensing data available for \( N=123 \) participants up until their graduation, consisting of 62 students from the 2021 cohort and 61 students from the 2022 cohort.
Section 2.4

Four Years of College Life

Figure 2.3: Temporal Progression of Self-Esteem: The figure shows participants’ self-esteem over the four-year study. The x-axis denotes the date, the y-axis the self-esteem score, and different colors represent different years. The smooth line indicates the average self-esteem score, with shaded regions displaying the 95% confidence intervals (calculated using bootstrapping). The data implies a gradual increase in participants’ self-esteem throughout the study.

Dartmouth College is located in a small college town in the United States and operates on a quarter system consisting of four ten-week terms: fall (mid-September to late November), winter (January to early March), spring (late March to early June), and summer (mid-June to early September). Throughout the academic year, undergraduates enjoy winter break, spring break, and summer break. Our study, initiated in 2017, enrolled two cohorts of students in a rolling fashion. Cohort 2021 (students graduating in Spring 2021) was enrolled in September 2017, followed by Cohort 2022 (students graduating in Spring 2022) in September 2018. We collected data from each cohort over four years, from their initial days on campus until graduation. We supplemented this data with regular self-reports on students’ mental health and well-being through self-reported EMAs. Figure 2.3 illustrates one such self-reported data
point (i.e., self-esteem) highlighting enrollment and study exit dates for both cohorts. It is evident that students experience a sudden drop in self-esteem upon joining the university, likely due to the challenging transition from high school to college-level academics, peer pressure, or homesickness. However, over the years, their self-esteem gradually increases, reaching higher levels by the time they graduate. In addition to EMAs and mobile sensing data, we administered several surveys at various time points during the study (every six months to a year). This extensive longitudinal study provided valuable insights into students’ behavior and mental health throughout their entire college experience.

Figure 2.4: Campus Events and PHQ4 Scores: The figure shows participants’ PHQ4 scores throughout the year, highlighting three significant campus events. The x-axis represents time, and the y-axis the PHQ4 score. Circular, diamond, and square markers denote different festivals: Homecoming (fall), Winter Carnival (winter), and Green Key (spring), respectively. A trend of lower PHQ4 scores around these events may suggest their calming effect on students’ mental health, with lower scores indicating less anxiety and depression.

Each term is characterized by a unique array of events and festivals that reflect the changing seasons and academic milestones. Fall term is characterized by *Homecoming*, a time when students, alumni, and faculty unite to honor the college’s rich history and traditions through various engaging events. Winter term brings the eagerly an-
2.4 Four Years of College Life  Capturing the College Experience

ticipated Winter Carnival, a weekend-long tradition that assembles the community to celebrate and enjoy the snow-filled season. As spring term arrives, the campus blossoms with vibrant flora and warmer weather, setting the stage for Green Key, an annual spring event featuring a diverse array of social, cultural, and athletic activities. These events are illustrated in our data presented in Figure 2.4. The x-axis represents the time period, while the y-axis displays the PHQ4 score. As detailed in Table 2.3 the PHQ4 (Patient Health Questionnaire-4) is a brief four-item questionnaire that students self-report once per week that is used to screen for depression and anxiety. The PHQ4 is based on the PHQ9 and the Generalized Anxiety Disorder-7 (GAD7). The PHQ4 has two subscales: the PHQ2 for depression and the GAD2 for anxiety. Each item is scored on a 4-point Likert scale that ranges from 0 (not at all) to 3 (nearly every day). The total PHQ4 score ranges from 0 to 12, and total PHQ2 and GAD2 scores range from 0 to 6. Higher scores denote greater levels of depression and anxiety. Results shown in Figure 4 suggest that these campus-wide festive events typically have a soothing effect on students, as their self-reported PHQ scores tend to be lower around the time of these events. There are some periods where the PHQ4 score, although not at its peak, is trending high. Especially in the year 2019 starting with Homecoming and a few events in 2018 as well.

The interviewed students suggested that the workload during these events could be the cause of the stress, as students are torn between enjoying the festivities or completing their work. The interviewed students pointed out that they often used to spend the weekends before these events doing their work to enjoy the festivities fully. Some students also shared that the increasing trend in PHQ4 in certain cases might be due to ‘frat ban’. First year students are not allowed to take part in fraternities and sororities of the college. Homecoming of the next fall (i.e., sophomore year) is when they are allowed to go into Greek spaces. Note, all partying/drinking events
2.4 Four Years of College Life  

Figure 2.5: Sleep Initiation Patterns: The figure shows students’ typical sleep initiation times, with different terms in various colors. Rectangles at the bottom indicate participant counts with sleep data per term. Summers feature earlier bedtimes, diminishing as the year progresses. The x-axis represents the time period, and the y-axis denotes sleep start time.

happens only at the Greek spaces on campus. So a few students shared that until that point, it is just first year students hanging out with each other and having closed groups. “But then after Homecoming, students have more access to the whole campus and that is when you start feeling like excluded from things because certain people go to certain places and not everyone can join. I think the consensus around the school is that Greek space is exclusive sometimes. So people start getting spread out and they might feel left out when others are doing something, increasing their stress.”, said one of the students we interviewed. Another student added a different perspective, “I am in a sorority and once I joined, I felt like it is nice to have things planned out for you, especially if you are a first year and you do not know where to go because it is your first festival, but then that kind of adds stress as I feel like I should be having fun because everyone tells me that it is fun”. Some students offered a different perspective on the impact of ‘frat ban’ lifting. They mentioned an increase in overall mood and a sense of becoming a full-fledged member of the campus community, which contributed positively to their college experience. They reflected that the lifting of the ban often
Four Years of College Life: Capturing the College Experience

coincided with the onset of exam periods, potentially complicating the interpretation of the observed spike in stress levels. Interviewed students also noted that while these events may change mental health at the moment, they may not have as much of an impact in the long run once students have experienced them.

In this extensive longitudinal study, we not only capture self-reported metrics from the students but also collect their behavioral data using our mobile app. Figure 2.5 illustrates one behavior we examined: the sleep start times of students across four academic years, specifically during the fall, winter, spring, and summer terms. The plot shows that sleep start times frequently shift within and across terms. Changes within a term may be influenced by significant milestones, such as mid-term and final exams. Across terms, we consistently observe that students tend to go to bed earlier during the summer, particularly in the first summer. This trend may be attributed to it being their initial summer break when many students are not yet engaged in

![Figure 2.5](image1.png)

(a) Winter 2018

(b) Fall 2019

(c) Spring 2020

Figure 2.6: Student Anxiety Patterns: The figure displays students’ anxiety levels across three terms: (a) Winter 2018, (b) Fall 2019, and (c) Spring 2020, highlighting mid-term weeks and final week as critical academic periods with heightened anxiety. The x-axis indicates the time period, and the y-axis represents the anxiety score.
2.4 Four Years of College Life  Capturing the College Experience

internships or other activities. During an interview, one student mentioned noticing a more balanced sleep schedule, especially during the summer break, attributing it to being at home and having fewer obligations to attend parties or socialize as they did on campus. However, the difference in sleep start times diminishes in the summer of 2019. At Dartmouth, nearly all second-year (sophomore) students are required to enroll and remain on campus during the summer term, participating in various academic, social, and outdoor activities (the second-year summer term is referred to as sophomore summer throughout the chapter). This requirement could contribute to the narrowing gap in sleep start times during this period. Additionally, the plot shows peaks in sleep start times towards the end and sometimes the middle of terms, potentially indicating mid-term and final exam periods. These peaks underscore the fluctuations in students’ sleep patterns as they navigate the academic calendar and its associated demands. Note that the numbers at the bottom of the plot indicate the number of students for whom data was available during the respective terms.

One important factor to consider in our study is Dartmouth’s quarter system, which offers students the flexibility to take terms off. This flexibility impacted our study, as some students went on study abroad programs or returned home, often choosing to disable the app during these periods. Participation in the study was entirely voluntary, allowing them the freedom to do so. Additionally, we observed a decrease in participation following 2020, which could potentially be attributed to circumstances related to the COVID-19 pandemic.

In Figure 2.6, we show the anxiety level of students in three different terms: Winter 2018, Fall 2019 and Spring 2020, with the exam periods highlighted. We observe that throughout the academic terms, students’ self-reported anxiety levels fluctuated, with noticeable increases typically occurring during midterm and final exam periods. It is evident that anxiety levels subside once the examination period
is over, allowing students to experience temporary relief from academic stress. As Dartmouth does not follow a conventional semester system, students do not have a fixed midterm examination schedule. Instead, they face the unpredictable academic schedule of midterms appearing at any time, ranging from as early as Week 3 to as late as Week 7. Consequently, the other peaks in the plot could also be attributed to this scattered exam schedule. However, midterms most often transpire during Weeks 6 or 7, which is why we have chosen to highlight these weeks in the plot. One of the students we interviewed stated that they were more stressed during midterms than finals, especially as a first year. They mentioned, “That was one of my first college exams, and I was not really sure what to think and I would be worried that if I bomb this test, there goes my grade. By the finals period, I knew what my grade was like and I would be less worried.” The final exam period typically takes place during the

(a) Cohort 2021 in Summer break 2018
(b) Cohort 2022 in Summer break 2019
(c) Cohort 2021 in Sophomore summer 2019

Figure 2.7: Student Stress in Summer: The figure shows stress levels of student cohorts during summers (a) Cohort 2021 in 2018 (summer break), (b) Cohort 2022 in 2019 (summer break), and (c) Cohort 2021 in 2019 (sophomore summer term i.e., second year summer). Separate cohort analysis ensures accurate representation, as second-year students stay on campus. The shaded area reveals increased stress towards summer’s end and fall term onset. The x-axis represents the time period, and the y-axis indicates the stress score.
last two weeks of the term, leading to a heightened level of anxiety among students as they prepare for and complete their final assessments. In addition, some final exams may be in the form of projects, which require students to start their preparations early, leading to increased anxiety in the lead-up to the final submission deadline. This early-onset stress can contribute to heightened anxiety levels throughout the term, even before the official final exam period begins. This pattern emphasizes the significant impact that exam-related stress, including both traditional exams and project-based assessments, can have on students’ mental well-being and underscores the importance of understanding and addressing the causes of anxiety throughout the academic term.

Figure 2.8: Partying/Social Patterns: The figure displays time spent by the 2021 Cohort at Greek spaces (fraternities and sororities) during their first two years. The x-axis indicates the date, and the y-axis represents the duration in Greek spaces, in hours. Different colors signify academic terms. First-year students are restricted from attending Greek spaces, lifting in their second year’s fall term. A peak in attendance occurs during the sophomore summer term, reflecting heightened social activity or prime party season.

Figure 2.7 illustrates student stress levels over two summers, segmented by cohort. The first two plots depict the first summer break for the 2021 (Summer 2018) and 2022 (Summer 2019) cohorts. In contrast, the lower figure, Figure 2.7c, highlights the increased stress during the sophomore summer term for the 2021 cohort, which
can be attributed to mandatory on-campus activities and classes (recall that second-year students have to stay on campus and enroll in classes during the summer). The highlighted regions in each graph indicate a surge in stress towards the end of summer, typically a few weeks prior to the Fall term commencement. This is an interesting phenomenon. This trend is particularly pronounced for the 2021 cohort in Summer 2018 (Figure 2.7a), where stress levels skyrocket near the summer break’s conclusion. The anticipation of returning to campus, facing academic pressures, and transitioning from a relaxed home environment to a more structured campus life, may contribute to this late summer stress spike. This observation was reinforced by our student interviews, where they noted a surge in commitments even before the official start of the fall classes. One student reflected, “Some sports and clubs have to begin preparing for the fall term a few weeks in advance. We also have to get ready packing for the rest of the year and preparing travel arrangements, which is stressful sometimes.” This suggests that both academic and logistical pressures are potential stressors as students approach the fall term. Several students highlighted that the summer break after the first year of college stands out from subsequent breaks. They pointed out that during the first summer break time, they often reconnect with high school friends who also return home. As the break ends, there’s a sense of leaving behind the comfort and familiarity of these long-standing friendships to return to the relatively newer college environment. This transition, they suggested, could contribute to the spike in stress levels experienced at this time.

Despite the escalating stress levels observed in the 2021 cohort throughout their sophomore summer, as shown in Figure 2.7c, this period paradoxically emerges as a peak party season for students. Figure 2.8 demonstrates a substantial surge in the time students spend in Greek spaces on campus during this second summer term. Sophomore summer is the biggest partying scene across all terms of college. This is
the summer when only the second year students are on campus. The teaching load is considered to be lighter than other terms. According to the students we interviewed, it’s typical during the sophomore summer term to enroll in only two courses, resulting in ample free time. Additionally, a campus-wide pong tournament (which often takes place in Greek spaces) is a popular summer event, with many students dedicating a significant amount of time to practice and participate. Students are clearly much more social as shown in the data. These Greek spaces, incorporating fraternities and sororities, serve as the primary venues for social events and drinking parties on campus. Note, all drinking parties are based in the Greek spaces on campus and are open to all students across campus. Thus, an extended time spent in these areas likely reflects enhanced engagement in social activities, essentially, more partying. However, a noteworthy observation from the figure is the negligible presence of students in Greek spaces until Fall 2018. This can be explained by a college-wide policy that bars first-year students, in our case, the 2021 cohort, from entering these Greek spaces (i.e., ‘frat ban’). Once this restriction is lifted after the first year, there is a considerable uptick in Greek space attendance during subsequent terms.

2.4.1. COVID Strikes

The COVID-19 pandemic struck midway through our study in March 2020, affecting Cohorts 2021 and 2022, set to graduate one and two years post-outbreak, respectively. Dartmouth College swiftly implemented CDC-recommended social distancing policies. Our application persisted in data collection throughout, enabling insights into the pandemic’s impact on student behavior and mental health. Furthermore, in order to enrich our analysis, we conducted interviews with students from the 2020-2021 and 2021-2022 academic years. These discussions, combined with the quantitative insights from the data, deepen our comprehension of the challenges students faced during this unprecedented time and underscore the significance of adaptability and
resilience in the educational landscape.

Figure 2.9: COVID-19 Concern Over Time: This figure charts the fluctuations in students’ concern regarding the COVID-19 pandemic over various time periods. The x-axis represents the time period, while the y-axis quantifies the COVID-19 concern scores. Different colors correspond to different years of the pandemic. The chart also accentuates several notable campus events. It appears that the level of COVID-19 concern among students responds to changes on campus, outbreaks, and the emergence of new variants (which is indicated in the plot with the vertical dashed lines).

Figure 2.9 shows students’ COVID-19 concerns from March 2020 to May 2022, with key time points corresponding to campus policy changes. Concerns peaked during the initial outbreak in early 2020, when remote teaching began. In May 2020, daily symptom screening started, and mask-wearing was enforced on campus. In September 2020, Dartmouth brought back more than half of the undergraduates, with safety measures like testing and face-covering policies, resulting in a rise in concern. In October 2020, a COVID-19 uptick occurred in the college town, but the on-campus community remained unaffected. Biweekly testing started in January 2021, and vaccinations became available in April 2021, causing a decline in student concern. However, the Delta variant in July-August and Omicron variant in December-January increased concerns. With vaccination rates rising and restrictions easing, concerns gradually decreased in 2022, and the university removed the indoor mask requirement by March.
Interviews with students who graduated in 2022 shed light on their experiences during the COVID-19 pandemic at the university. They reminisced about the summer of 2021, when a short-lived removal of the mask mandate was followed by a reinstatement due to a significant outbreak. By Fall 2021, in-person classes resumed without a mask mandate, offering students a sense of relief as the university seemed to be handling the pandemic effectively. Although some professors still held occasional Zoom sessions in Winter 2022, most classes took place in person. Dartmouth transitioned to remote instruction and pass/fail grading during the pandemic, providing students with greater flexibility in selecting their courses. This encouraged many students to take more challenging classes to improve their GPAs, leading to unique challenges for both students and professors as they adapted to online learning. However, the pass/fail grading system also induced stress for some students. One graduate shared their experience: “I think students were stressed up in the beginning and then we had pass/fail grading, so we laid up a little to the point where my friends and I sometimes, took a little easier on assignments. Professors mentioned that if we had a certain grade, we would not need to take the finals. And by the end of the term, during the finals period, we were just figuring out that we do not have the required grade. So we had to take the finals and also had to do well in the final in order to actually pass”. They added, “It was an ambiguous term and we did not know what accounted for pass or fail. We realized it would look a lot worse to have fail from a class rather than a B or a C”. While the majority of students appreciated the pass/fail grading system, some expressed dissatisfaction, believing that the lack of differentiation in grading reduced their incentive to work hard.

In summary, Dartmouth faced unprecedented challenges during the COVID-19 pandemic. By implementing policy changes, safety measures, and academic adjustments, it managed the crisis and started returning to normalcy. As the institution
moves beyond the pandemic era, it is important to observe the ongoing impact on student life and academic performance and apply the lessons learned to create a stronger, more adaptive educational environment for future generations. This chapter focuses on both pre-pandemic aspects and the significant influence of COVID-19 on students’ lives during the study period.

**Section 2.5**

**First vs. Final Term: Behavioral Shifts in Undergraduate Students**

The first term of college is a critical period during which students adapt to a new academic environment, establish social connections, and develop essential skills to help them succeed. In contrast, the final term is when students solidify decisions about future careers and/or further studies. Final-term students likely experience substantial growth in their values, beliefs, and sense of self, which may be evident in the behavioral data and self-reports we obtain from them. In this section, we compare the first term on campus, upon entering college as undergraduates, with the last term of their college experience, separately, based on their cohorts. Our data is non-normally distributed, which led us to choose non-parametric tests for our analysis. Specifically, we use the Wilcoxon signed-rank test, a non-parametric version of the paired t-test. This test is more appropriate for our data since we are comparing the differences in behavior between the first term and final term of the same students. We report the behaviors that are statistically significant after correcting for multiple comparisons using the False Discovery Rate (FDR) approach. FDR is a statistical measure that controls the proportion of false positives among statistically significant results in multiple comparisons testing. We list the result of comparison in Table 2.4.

We observed that Cohort 2021 participants (students who graduated in Spring
Table 2.4: First Term vs. Final Term: The table below presents the results from a Wilcoxon signed-rank test comparison between the first and final terms of two cohorts - Cohort 2020 and Cohort 2021. The significance levels, adjusted using the Benjamini/Hochberg FDR correction, are denoted as follows: (*** $p < .01$, ** $0.01 \leq p < .05$, * $0.05 \leq p \leq .10$). The ‘first’ and ‘final’ term columns represent the median scores for each period, while ‘W’ refers to the W-value, which represents the expected proportion of false positives among the statistically significant results. The epochs (0-3) correspond to different periods within a day: epoch 0 is the 24-hour average, epoch 1 is morning (12 am - 9 am), epoch 2 is daytime (9 am - 6 pm), and epoch 3 is nighttime (6 pm - 12 am). Note: All duration-related features, except ‘total time spent at home,’ are measured in minutes. The ‘total time spent at home’ feature is quantified in hours, while distance-based features are recorded in miles.

<table>
<thead>
<tr>
<th>Behavior/Self-report</th>
<th>Cohort 2021</th>
<th>Cohort 2022</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First term</td>
<td>Final term</td>
</tr>
<tr>
<td>Duration in vehicle [Epoch 0]</td>
<td>20.12</td>
<td>29.92</td>
</tr>
<tr>
<td>Duration in vehicle [Epoch 2]</td>
<td>11.35</td>
<td>19.40</td>
</tr>
<tr>
<td>Duration in vehicle [Epoch 3]</td>
<td>5.75</td>
<td>8.25</td>
</tr>
<tr>
<td>Still duration [Epoch 2]</td>
<td>415.56</td>
<td>431.64</td>
</tr>
<tr>
<td>Walking duration [Epoch 2]</td>
<td>84.28</td>
<td>9.02</td>
</tr>
<tr>
<td>Total time spent at home</td>
<td>10.68</td>
<td>11.39</td>
</tr>
<tr>
<td>Distance travelled [Epoch 0]</td>
<td>6.24</td>
<td>12.62</td>
</tr>
<tr>
<td>Distance travelled [Epoch 3]</td>
<td>27.86</td>
<td>27.86</td>
</tr>
<tr>
<td>Phone unlock duration per hour spent at home</td>
<td>6.97</td>
<td>9.75</td>
</tr>
<tr>
<td>Number of phone unlocks per hour spent at home</td>
<td>2.44</td>
<td>3.14</td>
</tr>
<tr>
<td>Max distance from campus [Epoch 3]</td>
<td>32.33</td>
<td>8.14</td>
</tr>
<tr>
<td>Number of locations visited [Epoch 0]</td>
<td>4.96</td>
<td>2.05</td>
</tr>
<tr>
<td>Audio play duration [Epoch 0]</td>
<td>114.90</td>
<td>161.64</td>
</tr>
<tr>
<td>Audio play duration [Epoch 1]</td>
<td>12.00</td>
<td>25.15</td>
</tr>
<tr>
<td>Audio play duration [Epoch 2]</td>
<td>51.50</td>
<td>79.38</td>
</tr>
<tr>
<td>Audio play duration [Epoch 3]</td>
<td>35.76</td>
<td>50.20</td>
</tr>
<tr>
<td>Number of audio plays [Epoch 0]</td>
<td>4.62</td>
<td>6.62</td>
</tr>
<tr>
<td>Phone unlock duration [Epoch 0]</td>
<td>139.79</td>
<td>212.56</td>
</tr>
<tr>
<td>Phone unlock duration [Epoch 1]</td>
<td>67.07</td>
<td>106.32</td>
</tr>
<tr>
<td>Phone unlock duration [Epoch 2]</td>
<td>49.74</td>
<td>66.51</td>
</tr>
<tr>
<td>Number of phone unlocks [Epoch 0]</td>
<td>69.27</td>
<td>91.58</td>
</tr>
<tr>
<td>Number of phone unlocks [Epoch 2]</td>
<td>33.30</td>
<td>47.72</td>
</tr>
<tr>
<td>Number of phone unlocks [Epoch 3]</td>
<td>26.80</td>
<td>31.76</td>
</tr>
<tr>
<td>Number of steps [Epoch 0]</td>
<td>8844.01</td>
<td>6773.00</td>
</tr>
<tr>
<td>Number of steps [Epoch 1]</td>
<td>1091.00</td>
<td>658.83</td>
</tr>
<tr>
<td>Number of steps [Epoch 2]</td>
<td>4820.68</td>
<td>3674.35</td>
</tr>
<tr>
<td>Number of steps [Epoch 3]</td>
<td>2061.64</td>
<td>2047.74</td>
</tr>
<tr>
<td>Self esteem</td>
<td>8.22</td>
<td>8.85</td>
</tr>
<tr>
<td>PHQ-4</td>
<td>1.78</td>
<td>2.08</td>
</tr>
<tr>
<td>Depression</td>
<td>0.50</td>
<td>0.92</td>
</tr>
</tbody>
</table>

2021) spent more time in their vehicles during their final term, both overall and during the daytime periods. However, they were less active during the day in their final term compared to their first term. They traveled less overall and during the night, which may explain the decrease in unique locations they visited. They used their phones
more frequently at home. Nevertheless, we found that they had a higher number of instances where they played media or audio, both overall and in the morning, day, and night. This trend was also observed in the number of audio plays. Additionally, both phone usage duration and the number of unlocks increased during the final term in all periods, except in the morning. We also noted that Cohort 2021 students took fewer steps during their last term compared to their first term. We observed similar findings for Cohort 2022 participants (i.e., students graduating in Spring 2022) in terms of vehicle usage—it was higher during the final term compared to their first term. They also had an increase in walking duration during the day and, unlike Cohort 2021, they were less sedentary in the day during the final term. They also spent less time at home or in their dorms. They remained closer to the campus during the night and visited fewer unique locations overall. In terms of self-reports, we found only three statistically significant variables: self-esteem, PHQ-4 score and depression score. Both cohorts had higher self-esteem in their final term compared to the first term. However, only Cohort 2021 had statistically significant difference in the first term and final term for PHQ-4 and depression score – both of which increased in their final term compared to their first term.

According to the interviews, the observed difference in behavior of the two cohorts might be attributed to COVID-19. The final term for Cohort 2021 was the spring of 2021 when the campus was still strict on gatherings and campus events, whereas for Cohort 2022, the final term was spring of 2022 when many COVID restrictions had been lifted. A lot of classes were online, and several students had pushed their major culminating experiences to spring, which the interviewees see as one of the reasons for the higher sedentariness in Cohort 2021 – they might have been just working more. They might have pushed off some of their hard classes and might have been doing their senior thesis along with it. Normally, the students explain, the last term is an
easy term. One of the interviewed students said, “in the last term, a bunch of people that we know, including us, are only taking two classes. I have so much time, I have more time than I have ever had. I do no work.” Another said, “a lot of people I know are just spending a lot of time hanging out with each other, especially since the weather gets nice.” This supports our findings from quantitative analysis for Cohort 2022, where they are more active than their first term.

In addition, the interviewed students mentioned that there is an annual college festival (i.e. Green Key) that takes place in week 7 of the spring term. Due to COVID, the festival was canceled in 2021 but it took place in 2022. The students interviewed mention that this could add to why walking duration is higher for Cohort 2022. Similarly, the higher phone usage duration for Cohort 2021 could be because of remote classes as well as to connect with friends. The proximity to campus, the students mention, could be because students get nostalgic as it is their final term. They know that they are graduating and they want to see all their friends. In addition, following the week when the final exams take place, there is a dedicated senior week when the seniors hang out with their class before the graduation ceremony. So students stay closer to the campus and leave late after graduation. In terms of vehicle usage, students noted that the duration in vehicle increased for the final term compared to their first term, probably due to students not being allowed to bring cars to campus in their first year.

Section 2.6

Gender Based Differences in Behaviors

Our study comprises 57 students identifying as male and 129 as female. We aggregated data for each gender over the four-year period and compared them using the Kruskal-Wallis test. We chose the Kruskal-Wallis test to compare behavioral differ-
2.6 Gender Based Differences

Capturing the College Experience

ences between the two independent groups, as our data was non-normally distributed. This non-parametric method avoids biases and inaccuracies from parametric tests that assume normality, ensuring a more accurate and robust analysis of our findings. Note that to account for multiple comparisons, we applied the False Discovery Rate (FDR) correction to the obtained p-values. The result is listed in Table 2.5.

Our analysis revealed several notable gender-based differences. Consistent with previous research (132; 177), male students exhibited higher levels of physical activity than their female counterparts. They walked more during all periods (overall, day, morning, and night) and recorded a higher number of steps (during the morning and night) than their female counterparts. In contrast, female students displayed higher levels of sedentary behavior. They had longer overall still durations across all periods, as well as longer still durations at social places and their own dorms. Our findings align with previous studies that have reported similar trends in gender-based sedentary behavior (323; 145). Interestingly, our analysis also revealed differences in audio play habits and sleep patterns. Male students listened to more audio (or played more media) on their phones, both in terms of frequency and total duration (overall and during the morning). On the other hand, female students reported longer sleep durations, as well as earlier bedtimes and wake-up times compared to males. Finally, we found that female students unlocked their phones more often during the morning (epoch 1).

In terms of the students’ self-reported surveys (i.e., EMAs), female students appeared more sociable, reporting that they spent more time with others than being alone (i.e., ‘social level’ question on the EMA), while male students reported higher self-esteem scores. Prior studies have shown that (based on how self-esteem is conceptualized), females tend to report lower self-esteem than males (161; 357). Regarding the EMAs conducted during the COVID-19 pandemic, female students reported
Table 2.5: Female vs. Male Participants: The table below illustrates the results from a Kruskal-Wallis test comparison between female and male participants in our study. The significance levels are adjusted using the Benjamini/Hochberg FDR correction and are denoted as: (*** $p < .01$, ** $0.01 \leq p < 0.05$, * $0.05 \leq p \leq 0.10$). The “Female” and “Male” columns display the median scores for each gender, and ‘H’ refers to the H statistic, which is a measure of the difference in median output scores between the two genders. A larger H statistic indicates a greater difference in median output scores. The epochs (0-3) delineate different periods within a day: epoch 0 signifies the 24-hour average, epoch 1 represents morning (12 am - 9 am), epoch 2 corresponds to daytime (9 am - 6 pm), and epoch 3 captures nighttime (6 pm - 12 am). Please note: All duration values, barring sleep duration, are expressed in minutes. Sleep duration is specified in hours. Both sleep start and end times are indicated in 24-hour format.

<table>
<thead>
<tr>
<th>Behavior/Self-report</th>
<th>Female</th>
<th>Male</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking duration [Epoch 0]</td>
<td>162.19</td>
<td>222.16</td>
<td>23.15***</td>
</tr>
<tr>
<td>Walking duration [Epoch 1]</td>
<td>17.51</td>
<td>26.05</td>
<td>22.79***</td>
</tr>
<tr>
<td>Walking duration [Epoch 2]</td>
<td>88.85</td>
<td>107.27</td>
<td>9.85***</td>
</tr>
<tr>
<td>Walking duration [Epoch 3]</td>
<td>56.33</td>
<td>79.29</td>
<td>27.70***</td>
</tr>
<tr>
<td>Still duration per hour spent at dorm</td>
<td>56.10</td>
<td>54.14</td>
<td>37.40***</td>
</tr>
<tr>
<td>Still duration per hour spent in social places</td>
<td>45.11</td>
<td>39.98</td>
<td>37.90***</td>
</tr>
<tr>
<td>Sleep duration [Epoch 0]</td>
<td>1222.15</td>
<td>1160.98</td>
<td>52.66***</td>
</tr>
<tr>
<td>Sleep duration [Epoch 1]</td>
<td>516.86</td>
<td>504.99</td>
<td>53.85***</td>
</tr>
<tr>
<td>Sleep duration [Epoch 3]</td>
<td>290.38</td>
<td>261.37</td>
<td>66.36***</td>
</tr>
<tr>
<td>Audio play duration [Epoch 0]</td>
<td>154.39</td>
<td>191.30</td>
<td>6.53**</td>
</tr>
<tr>
<td>Audio play duration [Epoch 1]</td>
<td>25.53</td>
<td>38.59</td>
<td>9.85***</td>
</tr>
<tr>
<td>Number of audio plays [Epoch 0]</td>
<td>5.75</td>
<td>6.85</td>
<td>5.55**</td>
</tr>
<tr>
<td>Number of phone unlock [Epoch 0]</td>
<td>9.73</td>
<td>10.96</td>
<td>6.53**</td>
</tr>
<tr>
<td>Sleep duration</td>
<td>7.62</td>
<td>7.19</td>
<td>5.17*</td>
</tr>
<tr>
<td>Sleep start time</td>
<td>23:48</td>
<td>00:27</td>
<td>15.56***</td>
</tr>
<tr>
<td>Sleep end</td>
<td>07:25</td>
<td>07:42</td>
<td>7.76**</td>
</tr>
<tr>
<td>Social level</td>
<td>3.23</td>
<td>3.02</td>
<td>10.36***</td>
</tr>
<tr>
<td>Self esteem</td>
<td>8.11</td>
<td>8.95</td>
<td>21.17***</td>
</tr>
<tr>
<td>COVID concern</td>
<td>4.16</td>
<td>3.66</td>
<td>7.76*</td>
</tr>
</tbody>
</table>

higher overall concern about COVID-19 and higher concern for themselves than their male counterparts. One of the female students we interviewed, shared her experience, “I think that sometimes college guys are a little more negligent, if that makes sense. A bit more self-absorbed. They act like it is their world that we are just living in. Whenever there was a COVID outbreak, all of my female friends would get tested whereas
the guys would always kind of go, I don’t know, whatever. I actually had COVID and
exposed some friends of mine. This was during my junior fall. All the girls that I told
were exposed, tested as they were worried. They were continually checking up to make
sure that they did not have it and take it home to their family because it was at the
end of the term. But all the guys did not care and they went home without a worry
in the world. So I feel like that could have been the COVID concern, even though the
girls might have been hanging out with each other more, they are more aware of the
impact whereas the guys might not think about it.”

While the students we interviewed could not provide any insights into the underly-
ing reasons for other gender-based differences, our findings, particularly those related
to physical activity, suggest that female students in our cohort may be at a higher
risk for negative health outcomes associated with sedentary behavior. As for audio
play habits and sleep patterns, the available research is inconclusive. Some studies
report that female students tend to sleep longer on free days (257), while others find
no difference in sleep duration (244). There is also limited research on whether female
students tend to unlock their phones more often than male students. Overall, further
research is needed to fully understand gender differences in audio listening habits,
sleep, and phone usage.

Section 2.7

The impact of changing living conditions

As the COVID-19 pandemic unfolded, universities had to adapt to remote learning.
In this section, we aim to investigate the effects of on-campus versus off-campus
living arrangements on students’ behaviors and mental health during the pandemic.
Dartmouth College began returning a limited number of students to campus in Fall
2020, with all students returning by Fall 2021. Consequently, we have three terms
of data for comparison: Fall 2020, Winter 2021, and Spring 2021. Note that for this analysis, we use the students’ GPS data to identify whether they are off-campus or on-campus. If they are within the 5 miles (ca. 8 km) radius to the campus, we term them to be on-campus, otherwise they are off-campus.

Figure 2.10: Propensity Score Matching: Figure (a) shows the treatment and control groups in comparison to one another post-matching. The closer the two distributions align, the more effective our matching procedure. Figure (b) presents the effect size of various covariates (calculated using Cohen’s D) pre and post-matching. Cohen’s D is a standardized measure of the difference between two means, expressed in standard deviation units. A decreased effect size post-matching indicates improved matching quality. While some effect sizes may increase post-matching, most of them remain below our threshold of 0.20. Both plots are specifically based on Fall 2020 data.

To account for potential confounding factors and non-random assignment of living arrangements, we employed propensity score matching (PSM) (264). PSM is a statistical technique used to create comparable groups of individuals when there is no random assignment to a treatment (in this case, being on or off campus). It aims to balance the distribution of observed covariates between treatment (on-campus) and control (off-campus) groups, thus reducing bias in estimating the treatment effect. In our case, the observed covariates include race, gender, cohort, socio-economic status (collected during their enrollment), and average PHQ4 score (which is reflective of their mental health). To perform PSM, we first estimate the propensity logit scores...
for each student based on the set of observed covariates using logistic regression. The propensity score represents the probability of being in the treatment group, given the observed covariates. Next, we match the students in the treatment and control groups based on their propensity scores, creating a matched sample with a similar distribution of covariates in both groups. We use nearest neighbor to create a matched sample with one-to-one matching. By matching on-campus and off-campus students based on the observed covariates, we created comparable groups for analysis. We performed PSM separately for each term and assessed the balance of confounding variables using standardized mean differences (SMDs). We show the plot of SMD of the covariates before and after matching in Figure 2.10b and the distribution of the on-campus and off-campus group after matching in Figure 2.10a. We iteratively refined the matching approach until satisfactory balance was achieved, as evidenced by most SMD below 0.20 in the plot. As shown in the plot, SMDs are reduced after matching, indicating that the covariate distributions are more similar between the treated and control groups. In other words, this ensures that the matching process has successfully minimized the influence of confounding factors. After matching, we used the Kruskal-Wallis test to compare behaviors and mental health self-reports between on-campus and off-campus students. If a significant difference was found, we calculated the average treatment effect on the treated (ATT) for each outcome to evaluate the impact of living arrangements during the pandemic. The result is listed in Table 2.6.

Our findings show a general trend of significant behavioral differences between on-campus and off-campus students during Fall 2020, which diminishes over time. For fall term, our matched sample has 52 participants each in on-campus (treatment) and off-campus (control) group. In Fall 2020, on-campus students visited more locations, engaged in more audio/media play, and unlocked their phones more frequently com-
pared to off-campus students. The ATT for still duration during morning (i.e., epoch 1) is approximately -8.97, indicating that the changes in living arrangement decreases still duration by 8.97 minutes on average for those students who returned to campus. Similarly, students on campus had an increase in walking duration during morning by 6.10 minutes on average. They also were using their phone more. In fact, on-campus students had on average, 40.08 minutes increase in phone unlock duration (overall). Being on campus also increased their phone unlock duration over the morning (epoch 1) and day (epoch 2). We observed no significant differences in self-reported mental health metrics. However, as the academic year progressed, the differences between on-campus and off-campus students became less significant. We have 42 and 21 students in each group during Winter 2021 and Spring 2021, respectively. However, by Winter 2021, fewer behavioral differences persisted, with on-campus students taking more steps overall and during different times of the day. Interestingly, we found that returning to campus during Winter 2021 decreased students’ social level by 0.42 units on average. The interviewed students pointed out why that might be the case, “There was the first outbreak in Winter 2021 and that was terrifying”. They added, “You could be put into isolation for a really long time even if you did not have COVID. You also could not get tested again. This is even if you were in just a general vicinity of the positive person. You were put into quarantine for two weeks. Everyone was afraid to contact trace anyone else in case they got mad at each other.” It is plausible that the outbreak could have led to the University imposing strict isolation, social distancing and testing guidelines, negatively impacting the social level of students. A few of the students mentioned that the increase in walking duration for on-campus students is probably because the only way they could have social time on-campus was by going outside. “And so a lot of people would go for walks and just be socially distant, like six feet apart, wearing your masks or whatever the restriction was at the
time”, mentioned one of the students. Students could not go inside buildings together, but they would just walk around campus a bit more so they could be social and talk to people without getting in trouble. Students also stated that the reduction in social level during winter might have also been caused by the weather itself which meant they could not be outside as much to go for their social walks.

Table 2.6: On-campus vs. Off-campus: The table below demonstrates the impact of changing living arrangements during the pandemic. We initially perform propensity score matching to pair the control group (off-campus) and the treatment group (on-campus), based on observed covariates. A Kruskal-Wallis test is then used to compare these groups. The significance levels, adjusted with the Benjamini/Hochberg FDR correction, are denoted as follows: (***) $p < .01$, (**) $0.01 \leq p < 0.05$, (*) $0.05 \leq p \leq 0.10$). The ‘Average Effect on the Treated’ (ATT) column reflects the average difference between the medians of the two groups. ‘Off-campus’ and ‘On-campus’ columns showcase the median scores for each group, while ‘H’ represents the H statistic. Epochs (0-3) correspond to different periods within a day: epoch 0 refers to the 24-hour average, epoch 1 to the morning (12 am - 9 am), epoch 2 to daytime (9 am - 6 pm), and epoch 3 to nighttime (6 pm - 12 am). Note: All durations, except sleep duration, are measured in minutes. Sleep duration is denoted in hours.

<table>
<thead>
<tr>
<th>Term</th>
<th>Behavior/Self-report</th>
<th>Off-campus</th>
<th>On-campus</th>
<th>H</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2020</td>
<td>Still duration [Epoch 1]</td>
<td>523.28</td>
<td>514.31</td>
<td>8.78**</td>
<td>-8.97</td>
</tr>
<tr>
<td></td>
<td>Walking duration [Epoch 1]</td>
<td>10.35</td>
<td>16.46</td>
<td>9.82***</td>
<td>6.10</td>
</tr>
<tr>
<td></td>
<td>Number of locations visited [Epoch 0]</td>
<td>1.57</td>
<td>2.01</td>
<td>15.75***</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Number of locations visited [Epoch 3]</td>
<td>1.22</td>
<td>1.45</td>
<td>12.30***</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Audio play duration [Epoch 0]</td>
<td>168.41</td>
<td>236.44</td>
<td>5.30**</td>
<td>68.03</td>
</tr>
<tr>
<td></td>
<td>Sleep duration</td>
<td>7.89</td>
<td>7.22</td>
<td>11.14***</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td>Number of phone unlocks [Epoch 0]</td>
<td>67.57</td>
<td>99.08</td>
<td>14.78***</td>
<td>31.51</td>
</tr>
<tr>
<td></td>
<td>Phone unlock duration [Epoch 0]</td>
<td>191.42</td>
<td>231.57</td>
<td>5.35*</td>
<td>40.08</td>
</tr>
<tr>
<td></td>
<td>Phone unlock duration [Epoch 1]</td>
<td>27.26</td>
<td>34.25</td>
<td>5.79**</td>
<td>7.00</td>
</tr>
<tr>
<td></td>
<td>Phone unlock duration [Epoch 2]</td>
<td>98.87</td>
<td>117.56</td>
<td>5.36*</td>
<td>20.70</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 0]</td>
<td>3317.39</td>
<td>5831.68</td>
<td>21.07***</td>
<td>2514.29</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 1]</td>
<td>273.13</td>
<td>545.14</td>
<td>21.92***</td>
<td>272.02</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 2]</td>
<td>2262.82</td>
<td>3824.18</td>
<td>19.19***</td>
<td>1561.36</td>
</tr>
<tr>
<td>Winter 2021</td>
<td>Social level</td>
<td>3.08</td>
<td>2.65</td>
<td>5.94*</td>
<td>-0.42</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 0]</td>
<td>3231.81</td>
<td>5188.76</td>
<td>14.61***</td>
<td>1956.94</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 1]</td>
<td>278.57</td>
<td>456.67</td>
<td>6.50*</td>
<td>178.10</td>
</tr>
<tr>
<td></td>
<td>Number of steps [Epoch 2]</td>
<td>2189.74</td>
<td>3477.32</td>
<td>15.00***</td>
<td>1287.58</td>
</tr>
<tr>
<td>Spring 2021</td>
<td>Number of locations visited [Epoch 3]</td>
<td>1.17</td>
<td>1.59</td>
<td>8.01*</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Moving on, in Spring 2021, only one variable remained significantly different between groups: the number of locations visited at night, with on-campus students visiting 0.42 more locations on average. We do not have enough participants off cam-
Section 2.8

The Lasting Impact of COVID-19 on Behavioral Patterns and Mental Health

In this section, we assess how the patterns in daily routines and mental health outcomes persist over time. In order to assess the behavioral changes over the four years, we group the students’ data into three periods: Pre-COVID (i.e., data from September, 2017 to March 13, 2020; beginning of the study to the date when COVID was declared a national emergency), During-COVID (i.e., data from March 13, 2020 to September 12, 2021 when classes were remote) and Return to campus (i.e., data from September 13, 2021 to June 15, 2022 when the majority of the students were back in Dartmouth’s campus for in-person instruction). We want to see if identified patterns during the COVID period persist once in-person classes begin and students return to Dartmouth or if students exhibit significant changes in their daily routines and mental health outcomes. If the patterns persist, this might suggest that the behavioral or mental health patterns identified during the COVID period have lasting effects on students. If the patterns change, it may indicate that students are adapting to the post-COVID period, and their daily routines and mental health outcomes are evolving. We used mixed effects modeling to account for the within-subject variability, repeated measures for each participant, and potential non-linear patterns in the
data. Our fitted mixed effects model, shown in Equation 2.1, considers each behavior or mental health outcome as a dependent variable. We use the participant identifier as a random effect and we control for gender, race, and the cohort the student belongs to. We use the period (i.e., pre-COVID, during-COVID and in-person instruction) as a fixed effect. This model will provide us with the average effects of each time period on the behavior measure and mental health outcome, controlling for gender, race and cohort, while accounting for the within-subject correlation through the random intercept for each student. We can then compare the estimated coefficients for the time periods to assess if there are significant differences in behaviors between the three time periods. Note that for this analysis, we only compare the epoch 0 (i.e., overall) behavior.

\[
Y_{ij} = \beta_0 + \beta_1 \cdot \text{Period}_{ij} + \beta_2 \cdot \text{Gender}_i + \beta_3 \cdot \text{Race}_i + \beta_4 \cdot \text{Cohort}_i + u_i + e_{ij} \tag{2.1}
\]

Here, \(Y_{ij}\) is the dependent variable (behavior or mental health outcome) for student \(i\) at time \(j\), \(\text{Period}_{ij}\) is a categorical variable representing the time period (pre-COVID, during-COVID, return to campus) for student \(i\) at time \(j\), \(\text{Gender}_i\), \(\text{Race}_i\), and \(\text{Cohort}_i\) are control variables for student \(i\), \(u_i\) is the random effect for student \(i\) (capturing the within-subject correlation), and \(e_{ij}\) is the residual error for student \(i\) at time \(j\). In this model, \(\beta_0\) is the intercept, \(\beta_1, \beta_2, \beta_3,\) and \(\beta_4\) are the fixed effect coefficients, and the \(u_i\) term represents the random intercepts for each student. We confirmed the assumptions for our mixed-effects analysis, including a Variance Inflation Factor (VIF) \(< 10\) for multicollinearity and assessing independence of effects. VIF is a measure of multicollinearity in statistical models, indicating how much the variance of an estimated coefficient is inflated due to collinearity. Note, most works consider VIF \(> 10\) as an indicator of multicollinearity. Despite non-normal data distribution, the mixed-effects model is justifiable given our large sample size and satisfied assumptions.
Data transformations and processing were performed to further ensure assumption adherence and enhance result reliability.

2.8.1. Passively Sensed Behavioral Patterns

The results for persistence of passively sensed behavioral patterns are listed in Table 2.7. Our findings show that the time spent walking decreased by approximately 52 minutes during the COVID period compared to the pre-pandemic level. However, after returning to campus, it increased by almost 3 minutes. The significant decrease in time spent walking during the COVID period compared to the pre-COVID period suggests that the pandemic had a negative impact on walking behavior. However, the increase in time spent walking after returning to campus compared to the pre-COVID period indicates that walking behavior is recovering or returning towards its pre-COVID baseline. Students were more stationary during the COVID period, with an increase in still duration of 62.81 minutes, and continued to have a higher still duration of about 9 minutes after they returned to campus. Similarly, students traveled less during the COVID period, with a decrease of 38.78 miles (ca. 62 km), and they continued to travel less after returning to campus (with an even higher decrease of 40.45 miles (ca. 65 km) compared to the during-COVID period, holding all other variables constant). Time spent on a vehicle decreased by 1.64 minutes during the COVID period compared to the pre-COVID duration, and it continued to decrease (with a higher rate of 3 minutes) even after the students returned to campus. Interestingly, one of the interviewed students’ perceptions seem to contradict our quantitative results, as they said, “I’m surprised by the vehicle finding, I was spending a lot more time in the car during COVID. I was driving because it was an activity, going to grocery stores more frequently.” But both of their observations in reduction in vehicle travel after returning to campus correspond with our findings, as they mentioned students were on campus and were not driving or traveling much.
Table 2.7: Behavioral Persistence: The table shows results from a mixed-effects model analyzing behavioral changes during COVID-19 and the transition back to in-person instruction compared to pre-pandemic times. Significance levels: (** p < .01, * .01 ≤ p < .05, .05 ≤ p ≤ .10). The ‘Pre-COVID’ period is the model’s intercept. Columns ‘β’, ‘S.E.’, and ‘Z’ display the coefficient, standard error, and Z statistics (which is a measure of how far a particular data point is from the mean of a dataset, expressed in standard deviation units) for each period, respectively, using only 24-hour average data (epochs 0). Note: Durations are in minutes, except sleep (hours), distances (miles), and sleep start/end times (24-hour format). Coefficients for changes in sleep start and end times during COVID-19 and upon returning to in-person instruction are in minutes.

<table>
<thead>
<tr>
<th>Behavior/Self-report</th>
<th>Pre-COVID (intercept)</th>
<th>During COVID</th>
<th>In-person instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>S.E.</td>
<td>Z</td>
</tr>
<tr>
<td>Duration in vehicle</td>
<td>29.00</td>
<td>2.41</td>
<td>12.00***</td>
</tr>
<tr>
<td>Still duration</td>
<td>1239.81</td>
<td>8.70</td>
<td>142.45***</td>
</tr>
<tr>
<td>Number of steps</td>
<td>7384.70</td>
<td>426.88</td>
<td>17.29***</td>
</tr>
<tr>
<td>Walking duration</td>
<td>133.18</td>
<td>9.92</td>
<td>13.43***</td>
</tr>
<tr>
<td>Distance travelled</td>
<td>116.06</td>
<td>19.44</td>
<td>5.97***</td>
</tr>
<tr>
<td>Sleep duration</td>
<td>7.33</td>
<td>0.19</td>
<td>37.03***</td>
</tr>
<tr>
<td>Sleep end</td>
<td>07:30</td>
<td>1.26</td>
<td>79.42***</td>
</tr>
<tr>
<td>Sleep start</td>
<td>00:01</td>
<td>1.38</td>
<td>30.35***</td>
</tr>
<tr>
<td>Unlock duration</td>
<td>181.00</td>
<td>11.18</td>
<td>16.19***</td>
</tr>
<tr>
<td>Number of unlocks</td>
<td>87.97</td>
<td>6.69</td>
<td>13.14***</td>
</tr>
<tr>
<td>Time spent at home</td>
<td>8.62</td>
<td>0.56</td>
<td>15.33***</td>
</tr>
<tr>
<td>Time spent at workout places</td>
<td>0.20</td>
<td>0.05</td>
<td>4.20***</td>
</tr>
<tr>
<td>Audio play duration</td>
<td>110.75</td>
<td>17.72</td>
<td>6.25***</td>
</tr>
<tr>
<td>Number of audio plays</td>
<td>4.09</td>
<td>0.44</td>
<td>9.30***</td>
</tr>
<tr>
<td>Time spent at study places</td>
<td>2.62</td>
<td>0.17</td>
<td>15.06***</td>
</tr>
<tr>
<td>Number of locations visited</td>
<td>3.74</td>
<td>0.14</td>
<td>27.28***</td>
</tr>
<tr>
<td>Time spent at social places</td>
<td>0.37</td>
<td>0.04</td>
<td>9.17***</td>
</tr>
<tr>
<td>Max distance from campus</td>
<td>1029.44</td>
<td>98.10</td>
<td>10.49***</td>
</tr>
<tr>
<td>Depression</td>
<td>1.03</td>
<td>0.17</td>
<td>6.09***</td>
</tr>
<tr>
<td>Anxiety</td>
<td>1.40</td>
<td>0.16</td>
<td>8.70***</td>
</tr>
<tr>
<td>PHQ-4</td>
<td>2.42</td>
<td>0.31</td>
<td>7.68***</td>
</tr>
<tr>
<td>Self-esteem</td>
<td>7.55</td>
<td>0.18</td>
<td>42.90***</td>
</tr>
<tr>
<td>Stress</td>
<td>2.58</td>
<td>0.09</td>
<td>26.83***</td>
</tr>
<tr>
<td>Social level</td>
<td>3.14</td>
<td>0.07</td>
<td>41.86***</td>
</tr>
<tr>
<td>COVID concern</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Our findings also demonstrate that the pandemic has influenced students’ sleep patterns and phone usage. During the COVID period, sleep duration increased by 0.168 hours compared to the pre-COVID period and continued to rise by 0.196 hours during the in-person instruction period, suggesting that students have maintained a longer sleep duration even after returning to campus. Sleep end hours increased by 0.530 hours during the COVID period, and this trend continued in the later period with an additional increase of 2.363 hours. In contrast, sleep start hours initially
decreased by 0.81 hours during the pandemic, but later increased by 0.79 hours in the post-COVID period, indicating a shift towards later bedtimes. Regarding phone usage, phone unlock duration increased by 33.85 minutes during the COVID period compared to the pre-COVID period, but only increased by 2.53 minutes after returning to in-person instruction, suggesting that phone usage is in the verge of returning to its pre-pandemic levels. A possible reason for increase in phone usage during COVID is to attend the remote classes. One of the interviewed students mentioned, “there was definitely a huge increase in the amount that I used my phone during COVID because I would be calling my friends for hours and I was just on Zoom or on Facetime with friends.” One of the students recalled that she would also frequently watch movies with her friends, “Netflix has an extension where you can watch a show with someone else from a different location. So once students were in their room, they probably were on their phones more. I would do that to socialize through phone.” The number of phone unlocks decreased by 17.06 units during the COVID period, and although this number increased slightly during the in-person phase, it still remained lower than the pre-COVID baseline with a decrease of 13.24 unlocks. Audio playing duration increased by 46.40 minutes during the COVID period, and this trend continued after return to in-person instruction, with an increase of 37.76 minutes compared to the pre-COVID level, indicating that students engaged more with audio content during the pandemic as well as after return to campus. The number of audio plays increased during the COVID period by 1.209 units and continued to increase in the latter period by 0.833 units, which supports the observation of increased engagement with audio content. The number of unique locations visited by students decreased by 2.641 during the COVID period, reflecting the impact of lockdowns and restricted movement. However, after returning to campus, the decrease was smaller at 1.975 unique places visited, suggesting that students started to visit more locations as restrictions eased.
2.8 Lasting Impact of COVID-19 Capturing the College Experience

The maximum distance traveled from campus decreased by 222.85 miles (ca. 358 km) after returning to campus compared to the pre-COVID period, which could be attributed to changes in students’ travel patterns or preferences after the pandemic subsides.

One of the students we interviewed found the reduction in steps even after return to in-person instruction interesting. She shared that it might be because a lot of classes had Zoom alternatives even though most of the students were back on campus – "Every class that was in person was still required to be recorded. So you could not be penalized for missing class. Whereas right now, if you are not there for attendance, you get marked off. Even though you were encouraged to go in person, the professors were also very lenient, encouraging kids to stay home if they did not feel well. I think classes in 2021 and the entire of 2022 were all recorded. So I know people sometimes will skip classes and be like I’ll watch the lecture later. A lot of kids had gotten used to not going to classes, so they were either saying I’m not feeling well to watch the recorded lecture or they did actually have COVID and I think that might have reduced the steps.” Another contributing factor, as suggested by a student, could be the growing number of personal vehicles among students in their later college years. This shift results in greater accessibility to rides among peers, reducing the need to depend on upperclassmen for transportation and consequently, possibly leading to a decrease in steps.

2.8.2. Self-reported Mental Health Outcomes

Our analysis of mental health outcomes reveal important changes in students’ mental health during the COVID period and after return to in-person instruction compared to the pre-COVID period. Persistence of self-reported outcomes is also listed in Table 2.7 towards the bottom. As shown in the table, depression scores increased by 0.430 during the COVID period and continued to be elevated by 0.334 after that.
Anxiety scores also increased during the COVID period by 0.322, with a slightly smaller increase of 0.239 after returning to campus. These findings may suggest that the pandemic had a negative impact on students’ mental health, with higher levels of depression and anxiety observed during and after the COVID period. The PHQ-4 scores, a measure of psychological distress, increased by 0.750 during the COVID period, indicating heightened distress among students. In the in-person period, the increase was smaller at 0.574, suggesting some improvement in psychological distress after the return to in-person instruction. Self-esteem scores increased marginally by 0.225 during the COVID period and continued to increase by 0.325 after that, indicating a slight improvement in students’ self-esteem throughout the pandemic and afterward. One of the interviewed students shared that they believed self-esteem would naturally improve as one got older. But the other mentioned that there could be two sides to it. As one grows older, it could be isolating as groups form and one may either feel a lot better about oneself or a lot worse, depending on whether they are finding their place. Both the students shed light on the impact of COVID in shaping their self-image, with one saying, “I think something that COVID did especially here was stratified friends and friend groups. If you did not have a really strong friend group since before the pandemic, then it was really hard for a lot of people to talk to others and feel like a friend group and that they belonged to the same college. People who did not have such groups would see all of their friends hang out without them and then when they came for in person, the groups who hung out during COVID would be so much closer because they were around. So they did not feel like they maybe fit into their groups as much.”. The other student supported this remark and mentioned, “There was a lot more cliques, friend groups were a lot smaller, and that probably either helped improve the self image or decrease it.”

Following that, stress scores increased by 0.260 during the COVID period and
remained elevated, with an increase of 0.229 after the period. This suggests that stu-
dents continued to experience higher levels of stress even after returning to campus. Social levels decreased by 0.326 during the COVID period, reflecting the impact of social distancing and lockdown measures on students’ social lives. Although social levels improved slightly after returning to in-person instruction, they remained lower than the pre-COVID baseline, with a decrease of 0.278. This indicates that students’ social lives have not yet fully recovered from the effects of the pandemic. COVID concern scores were only available for the during-COVID and in-person instruction periods. During the COVID period, the COVID concern score was 4.076, which decreased by 0.553 during the in-person instruction period. This suggests that students’ concerns about COVID-19 have diminished as the pandemic situation improved and they returned to campus. One of the students shared that, “I think the mental health issues will eventually baseline out a bit more, but it might take a longer time. Because I feel there is a subset of people who realized that they might have issues with depression and anxiety and their mental health that they did not really realize until they were locked in their home alone. And so I think that might eventually go back once they have more social interaction and other care to reduce those levels, but probably not super fast.”

In summary, our findings suggest that the COVID-19 pandemic might had a sig-
nificant impact on students’ behaviors and mental health outcomes. Although we do not have a control group to directly compare changes over time, the unique cir-
cumstances and timing of the pandemic provide strong indications of a potential association. Some behaviors, such as time spent walking and phone usage, improved or returned to pre-COVID levels after resuming in-person instruction, while others, like time spent in vehicles, sedentary duration, and distance traveled, continued to
show persistent changes. Mental health outcomes also appear to have been affected, with increased levels of depression, anxiety, stress, and psychological distress observed. While self-esteem improved during the decline of the pandemic and in-person instruction period, social engagement levels have not fully recovered. The decrease in COVID concern scores after returning to in-person instruction suggests that students’ worries about the pandemic have diminished as they adjusted to the new normal. Overall, the observed changes in students’ behaviors and mental health outcomes during the pandemic and after in-person instruction underscore the potential complex and multifaceted impact of COVID-19 on the college population. Monitoring these trends and providing appropriate support remains important as students continue to adapt and return to in-person instruction.

Section 2.9

Discussion

In this section, we discuss our findings, the implications of our work and the associated ethical considerations.

2.9.1. Summary of findings

Our data highlights the changes in daily routines and physical activity levels among students throughout their four years at Dartmouth. Particularly noteworthy was the comparison between first and final term students across different cohorts. Notably, the 2021 cohort displayed heightened sedentary behavior, which could be indicative of an increased workload and the postponement of challenging courses or thesis work. In contrast, the 2022 cohort demonstrated greater physical activity and less time spent at home, potentially reflecting the reintroduction of social events and a more vibrant campus life. We also observed gender differences, with male students generally being
more active while female students exhibited higher levels of sedentary behavior and expressed more concern about COVID-19. This highlights the possibility of differential health risks and underscores the need for gender-specific interventions (320; 216). Furthermore, we examined the contrasting experiences of students living on-campus versus those living off-campus during the pandemic. Our findings showed significant behavioral variations between these groups, reinforcing the importance of tailored support and resources to cater to these differing circumstances (13).

The ramifications of the COVID-19 pandemic on student behaviors, as discussed in our study, extend beyond the immediate impact of the crisis. As Dartmouth College pivoted back to in-person instruction, our research observed a persistence of certain behavioral patterns that were adopted during the remote learning period. This phenomenon was particularly evident in physical activity and technology usage. For instance, we noted a sustained decrease in distance travelled and vehicle use even after the resumption of in-person classes. Furthermore, the increased level of stillness during the pandemic period seemed to persist, indicating a possible long-term shift towards more sedentary behaviors. Another interesting observation was related to phone usage. While it was expected that phone usage would increase during the height of the pandemic due to remote learning and reduced face-to-face interactions, we discovered that even with the return to in-person instruction, phone usage did not revert entirely to pre-pandemic levels. This could reflect a lasting shift in communication and study habits, where students continue to rely heavily on digital tools despite the availability of in-person options.

These persistent behavioral changes underscore the profound and enduring impact of the pandemic on students’ lifestyles and habits. It also illustrates the adaptability of the student population in embracing new routines and practices in response to external circumstances (201). Moreover, these observations raise important con-
siderations for universities and policymakers in terms of providing support for these behavioral adjustments, such as enhancing infrastructure for active transportation and providing resources to manage screen time effectively. This persistence of behavioral change is also of significant interest to the HCI field. It underscores the need for adaptive technologies that can seamlessly transition between different contexts (in-person, remote, hybrid) and continue to provide support as users’ behaviors evolve (70). Future HCI research could explore ways to design and implement such flexible technologies, further contributing to the resilience and well-being of students in a post-pandemic world.

2.9.2. Implications

Our findings have important implications for universities, policymakers, and mental health practitioners, as they reveal key insights into the factors that contribute to students’ ability to adapt and thrive during challenging circumstances. By understanding the unique experiences and challenges faced by students during their college journey, we can better support their mental health and well-being, ensuring that they have the resources and tools they need to succeed academically and personally.

For the HCI field, our research highlights the potential of mobile sensing technology and innovative data collection methods in capturing fine-grained, real-time information on student behaviors and mental health, importantly, over long periods of time. This opens up new avenues for the development of context-aware applications and interventions that can support students in managing their well-being throughout their college years. Our study promotes interdisciplinary collaboration between the HCI field and other domains, such as psychology, education, and public health. By working together, experts from various disciplines can develop a more holistic understanding of the complex factors influencing college students’ mental health and well-being, leading to more effective and contextually appropriate solutions. Educational
institutions can benefit from our findings by implementing targeted support programs and policies that consider the unique challenges and needs of students during different stages of their college journey, including during times of crisis. This could involve fostering a supportive campus environment, offering mental health resources, and creating flexible academic structures that accommodate students’ diverse needs. For mental health professionals, our study provides evidence of the associations between daily activities, living arrangements, and mental health outcomes, emphasizing the importance of considering students’ behavioral patterns and environmental factors in the development of effective interventions and support strategies. This knowledge can inform the design of tailored, evidence-based approaches that promote resilience and well-being among college students. Many universities are augmenting student health services with a number of digital tools to support the mental health of students, such as, mindfulness apps and online access to mental health therapists. With the advance of AI and other digital tools there is likely to be a new wave of intervention technologies that could also integrate behavior health driven by mobile sensing. Lastly, our findings have implications for policymakers and public health officials, who can leverage this information to advocate for mental health initiatives and support services targeting college students at a national level. By understanding the broader impact of the pandemic and other environmental factors on students’ mental health, policymakers can develop informed strategies and allocate resources to ensure that the well-being of this population remains a priority.

The findings from our comprehensive four-year study can offer valuable insights to student advocacy groups, including Active Minds (219) and the Jed Foundation (105). Moreover, numerous colleges are actively sharing their annual institutional mental health data, contributing to the national Healthy Minds Study (235). The inclusion of validated measures from the Healthy Minds Study survey provides essential
2.9 Discussion

Capturing the College Experience

national data on the prevalence of mental health outcomes and attitudes about mental health on college campuses. Examining mental health across diverse campuses, both locally and nationally, is crucial for ensuring the reproducibility and generalizability of results, as demonstrated in this chapter. A collective effort involving additional studies at various types of educational institutions, both at the undergraduate and graduate levels, national survey outcomes, and advocacy by diverse groups will significantly contribute to advancing our understanding of student mental health. This collaborative approach aims to reduce stigma, enhance knowledge, and ultimately eliminate barriers to mental health service utilization, thereby addressing the mental health crisis prevalent on college campuses.

In summary, this study offers valuable insights and implications for various stakeholders involved in the college experience, highlighting the potential of interdisciplinary research and the importance of collaboration between the HCI field, educational institutions, mental health professionals, advocacy groups and policymakers to promote the mental health of college students.

2.9.3. Ethical Considerations

The design and execution of our study were underpinned by a firm commitment to ethical research practices. All participants were informed about the nature of the study, the type of data collected, and its intended use. Informed consent was obtained from all participants prior to the start of the study. The data collected was anonymized and securely stored, ensuring that individual participants could not be identified. Moreover, privacy was a paramount concern throughout the study, and we took steps to ensure that the data collected did not infringe upon the participants’ personal lives. We also respected participants’ right to withdraw from the study at any time without penalty. Furthermore, this study was approved by the Institutional Review Board (IRB) of the involved institution, affirming its adherence to ethical
research standards. It is important to note that the insights and conclusions drawn from this study are meant to contribute to the broader understanding of college experience, student’s behavior change due to COVID and student resilience and are not intended to label or stigmatize individuals or groups.

Section 2.10

Limitations and Future Work

Despite the valuable insights provided by this study, it is important to acknowledge its limitations, which may have implications for the interpretation and generalizability of the findings.

First, the sample size and demographic of the participants are limited to students from a single institution, which may not be representative of the broader college student population. Next, our analysis is based on self-reported mental health metrics, which may be susceptible to recall bias and social desirability bias. The use of objective measures, such as clinical interviews or assessments, would help to validate and augment the self-reported data. Another limitation is the potential for attrition bias. As the study spans four years, there may be differences between the students who remained in the study and those who dropped out or had incomplete data. These differences could lead to biased estimates of the effects under investigation. This study predominantly focuses on the associations between various behavioral factors and mental health outcomes. While these associations provide valuable insights, they do not necessarily imply causality. Unmeasured confounding factors or other contextual influences may be responsible for the observed associations, and future research should aim to explore the potential causal pathways underlying these relationships. This is also true for the comparison we did between on-campus and off-campus students. Although we used propensity score matching (PSM) to control for potential
confounding variables when comparing on-campus and off-campus students, we may not have accounted for all confounding factors. There might be other unobserved variables that could influence the differences observed between the two groups. Similarly, although we have conducted analyses to assess the presence of MCAR and MAR in our dataset, it is important to acknowledge that these analyses might not be entirely reliable due to the complex nature of missing data mechanisms. The methods we employed to detect MCAR and MAR, such as Little’s MCAR test and correlation analysis, have their own limitations and assumptions that may not hold in all situations. As a result, our conclusions regarding the presence of MCAR or MAR should be interpreted with caution. Furthermore, it is difficult to directly test for missing not at random (MNAR), as it depends on unobserved data. We acknowledge that our study might be subject to potential biases arising from MNAR, and our conclusions should be interpreted with caution. Given the limitations of our dataset, we cannot definitively rule out the possibility of MNAR. To account for the potential impact of MNAR, as well as the limitations of our MCAR and MAR analyses in future research, it would be beneficial to incorporate additional data sources or use data collection methods that minimize the likelihood of missing data related to unobserved variables. For example, researchers could use wearable devices or smartwatches in combination with smartphones to collect more comprehensive data on participants’ behaviors, or they could employ more frequent follow-ups and reminders to encourage participants to provide complete data. Related to this, there are potential biases that may arise due to differences in how participants carry their phones, which could impact our conclusions, such as the observed difference in sedentary behavior between male and female participants. The limitations in our data collection process, including the inability to directly account for how and where participants carry their phones, are important factors to consider when interpreting our findings. Incorporating additional
data sources, such as activity trackers or smartwatches could help to obtain a more accurate and comprehensive picture of participants’ behaviors and better account for potential biases related to phone usage and carrying habits.

Although our sample is limited to a single institution and may not represent the broader college student population, the findings still offer valuable insights into the experiences of students facing similar challenges during the pandemic. These insights can serve as a foundation for future research and contribute to the design of interventions promoting resilience and well-being among college students in various settings, including low and middle-income countries (LMICs). We recognize the potential limitations in generalizing our findings to LMICs, where students may encounter distinct challenges and access different resources. Nevertheless, the insights obtained from this study can benefit LMICs by providing an understanding of the associations between daily activities, living arrangements, and mental health outcomes during a global crisis. This knowledge can guide the design of interventions and support strategies tailored to the unique needs and circumstances of students in LMICs. To further enhance the generalizability of our findings, we recommend future research involve multi-institutional collaborations across various countries, including LMICs. This approach would enable researchers to capture a more diverse range of student experiences, resulting in a comprehensive understanding of the factors influencing college students’ mental health and well-being during challenging times. In addition, incorporating qualitative data by interviewing a larger and more diverse group of students would enrich the insights provided by this study. In this study, we employed multiple single-item measures, such as stress, social level, and COVID concern, which may not fully capture the complexity of these constructs. Our choice to use single-item measures was motivated by our desire to reduce participant burden and encourage higher response rates, given the repeated nature of our data collection. We were able
to find some evidence of criterion validity post-study by correlating these single-item scores with other mental health outcomes. For example, we examined the association between the single-item stress measure and the PHQ4, finding a strong positive correlation between stress and the total PHQ4 score ($\rho=0.64$), depression ($\rho=0.58$), and anxiety ($\rho=0.63$) subcomponents, which are known to be associated with stress. But still, we acknowledge that more comprehensive measures, such as validated multi-item scales, could provide a more accurate representation of students’ stress, social level, and COVID concern throughout the semester. Future research should consider using more extensive measures to assess these constructs and further validate the findings of this study.

Section 2.11 College Experience Study Dataset and Open Research Questions

In the interest of advancing research into the mental health of college students during their college years using mobile sensing and self-reports, we discuss the public release of the de-identified College Experience Study dataset on Kaggle (142). We also outline a series of open research questions that could be studied using the public dataset. We encourage the use of this dataset for academic and research purposes and request that any publications resulting from its use appropriately cite it as follows:

We hope that this contribution will facilitate a deeper understanding of the pandemic's impact on student life or the student life in general and promote further studies in this domain. There are many open questions that researchers can study with this dataset; for example

- **Student Behavior During National Events:** The longitudinal data spanning four years offers a unique opportunity to analyze student behavioral shifts in response to national events, such as electoral cycles. Research could investigate how significant societal moments correlate with changes in student mood, activity levels, and social interactions.

- **Cross-Institutional Comparisons:** The unique dataset allows for broader comparisons, leading to the question: How do the behavioral and mental health trends observed in this study compare with those at other institutions, both within the US and internationally? Our dataset can act as a benchmark for educational institutions to identify pervasive behavioral patterns and institution-specific trends, aiding in the creation of targeted mental health initiatives.

- **Model Pre-training and Transfer Learning:** Smaller studies often face constraints in data availability. Leveraging our extensive dataset for pre-training models could enhance their predictive performance through transfer learning, thereby optimizing research efficacy and reducing resource expenditure.

- **Longitudinal Mental Health Trajectories:** How do mental health trajectories evolve over the course of college life? The dataset allows for an examination
of long-term trends and the identification of critical periods for intervention.

- **Socioeconomic Status and Social Dynamics:** Investigating the influence of self-perceived socioeconomic status on students' social interactions might provide insights into social stratification within college environments.

## Conclusion

In conclusion, our longitudinal mobile sensing study provides a comprehensive analysis of undergraduate students’ behaviors, mental health metrics, and experiences throughout their four-year college journey at Dartmouth College, with a particular focus on the unprecedented challenges presented by the COVID-19 pandemic. Through the utilization of a mobile sensing app and supplementary qualitative insights from student interviews, we identified patterns and associations between daily activities, living arrangements, and mental health outcomes. This study not only offers valuable insights into the impact of the pandemic on college students but also emphasizes the significance of understanding the broader context of their experiences throughout their college years. The findings of this study contribute to the existing body of literature on the relationship between daily activities and mental health, highlighting the ongoing need for research and support systems that foster resilience and well-being among college students. The study also underscores the potential of mobile sensing technology and interdisciplinary approaches to illuminate the interplay between behavior, environment, and mental health within higher education settings. Overall, this study emphasizes the importance of a holistic understanding of college students’ experiences and provides a foundation for the development of targeted interventions and support mechanisms that address the unique challenges they face. Finally, we publicly released the College Experience dataset (142) to help accelerate research in
this field.
Chapter 3

Detecting Job Promotion in Information Workers Using Mobile Sensing

Section 3.1

Introduction

Promotions play an important role for organizations and individuals. For organizations, it is a way to keep employees committed and motivated towards the company goals by rewarding promoted workers with financial and status gains. For individuals, rising through the ranks leads to a boost in morale, wellbeing, and life satisfaction (149; 150). However, promotions can be a mixed blessing for many – while they provide an increase in occupational status, financial reward, job autonomy, privilege and flexibility, they can often also be accompanied by added responsibility, longer working hours, stress and reduced work-life balance. Promotions are therefore, a win-some, lose-some game. While workers may win through the status gain, financial and personal growth, they may impact their psychological wellbeing and work-life
balance. In addition, job change or role transitions have the ability to change people
(237; 7; 51; 167). The characteristics of a person may change as a result of change
in occupational status (109). As people transition in their new role after promotion,
it might involve psychological, cognitive and behavioral adjustments. Workers might
need to embrace their new role in the social network, learn new tasks, get familiar
with new routines as well as cope with potentially new physical settings and sur-
roundings (166; 83). As such, these transitions can have positive as well as negative
consequences on a promoted worker’s life. Therefore, we believe studying the af-
ter effects of promotion is an important topic. Prior studies have linked workplace
promotions with health, mental wellbeing, job satisfactions among other adjustment
issues for workers (46; 147; 103; 106).

Despite their importance, we do not have any objective assessment of how promo-
tion affects an individual’s life. In the absence of literature using objective measures
to capture the effects of promotion, we propose a passive sensing based approach as
a means to assess workers’ reactions to being promoted. We do not concern our-
selves with how the promotions happen or why they do, because the subject has been
studied thoroughly before (28; 343; 300; 104). Rather, we focus on physiological and
behavioral changes brought about as a result of promotion. More specifically, we look
at the period before promotion (i.e., the non-promoted period) and after promotion
(i.e., the promoted period) in order to examine differences. We study N=141 informa-
tion workers across different industries who were promoted during the year long
study using data from workers’ mobile phones, wearables and Bluetooth beacons.
We compare these physiological and behavioral changes before and after promotion
taking into account gender and job performance. Unlike prior studies that use self-
reported measures to assess the impact of workplace promotion (149; 150), we believe
that passively collected data from mobile devices can provide an objective measure
3.1 Introduction Detecting Job Promotion Using Mobile Sensing

of promotion events. The contributions of this paper are as follows:

(a) To the best of our knowledge, we present the first workplace based mobile sensing study to investigate the objective behavioral and physiological changes that occur as a result of promotion. We consider N=141 information workers who are promoted in a year long study. We do this by performing a comparison of workers’ physiological and behavioral patterns up to 60 days prior to promotion and 60 days after promotion.

(b) We study the changes in physiological and behavioral patterns associated with promotion considering gender and higher/lower job performers. Our findings show that there are gender differences associated with promotion events; for example, after promotion, female workers experience an increase in stress duration during working hours, but males do not.

(c) We explore the feasibility of using mobile sensing streams to detect job promotion events. Using a binary segmentation based changepoint detection method, we find that mobile sensing streams (e.g., sleep duration, step count) change abruptly during the promotion period, illustrating that the signals have the potential to capture promotion.

(d) We use ROCKET (86), a random convolutions based approach to extract patterns from multivariate time series signals. Using 1D convolutions, we extract features from the multivariate time series and evaluate the performance of different models to classify whether the given mobile sensing data belongs to a promoted period or a non-promoted period. For this analysis, we include non-promoted as well as promoted workers’ data. We report an AUC of 0.72. Our findings show that passively collected mobile sensing data from phones, wearables and beacons is able to detect job promotion events.
3.2 Related Work

The structure of the paper is as follows. We first describe the related work in Section 3.2 and then detail our study and data collection in Section 3.3. We discuss our analysis and results in Section 3.4 and Section 3.5, respectively. Following this, we discuss the implication, insights and possible applications of our work in Section 3.6. In Section 3.7, we discuss the important issue of safeguarding workers’ rights as new sensing technologies are developed in the future of work era (22). Finally, we discuss the limitations of our work in Section 3.8 and present some concluding remarks in Section 3.9.

Section 3.2

Related Work

The majority of prior research related to workplace promotion is focused on an individual’s likelihood for being promoted. This topic is studied in relation to both job performance and other related factors. For example, researchers find related factors, such as, personal characteristics, psychological attributes and education level are more related to being promoted than simply performance on the job (28; 343; 300; 104). Researchers use personality characteristics, job attributes and psychological information to train machine learning models to predict whether an employee is likely to be promoted. Long et al. (188) use demographic (e.g., gender, date of birth, etc.) and job features (position, department, position type etc.) to predict if an employee (N=71132) is promoted or not using a Random Forest model reporting an AUC of 0.96. Other research (355) reports a correlation between work related interactions (i.e., workplace blogs, assigning tasks, downloading work related files, etc.) and online social connections with employee promotion and retention. Work related interaction is strongly predictive and correlates with promotion and retention. The authors (355) collect data from an internal social network platform used by the company and train
a logistic regression model to predict promotion and resignation of N=104 employees.

There is little in the way of work examining transitions around promotion specifically pre-post promotion behavioral and physiological changes reported in this paper. Prior work investigating changes around job promotion purely use self-reports associated with general health, psychology, happiness and other metrics. Boyce et al. (46) use the British Household Panel Survey (BHPS) data collected annually over 16 years and report that promoted individuals suffer from deterioration of psychological wellbeing. Johnston et al. (147) use the Household, Income and Labour Dynamics in Australia (HILDA) survey across 9 years to analyze promotion effects. The authors (147) find that while promotion improves job security, pay perception and job satisfaction, there is negligible effect on general health and happiness. Similar to Boyce et al., the authors report that promotion negatively affects the mental health of workers. However, several studies show that working in low-ranked jobs also leads to poor mental wellbeing (103; 301; 122) and social functioning (302). Because promotion is often followed by a rise in job rank, the findings from these studies suggest that promotion may in fact lead to an improvement in overall health. Along the same lines, researchers also find that promotion may reduce the probability of heart disease by up to 13 percent over a period of 15 years (11). In a study of N=871 employees, de Lange et al. (83) find that after promotion, employees report an increase in job autonomy and work engagement. With respect to “adjustment” after promotion, Kramer et al. (166) show that employees need to create new communication relationships and skill sets to become comfortable in new roles. Other prior literature examines gender differences in behavior after promotion. Nyberg et al. (242) report that self-rated health decreases for both males and females after promotion. However, promoted females in the study report greater health decreases in comparison to males. Johnston et al. (147), on the other hand, report that promotion impacts the mental health of young
males more severely than that of females (147).

There is a growing interest in future of work research. The Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) community is well placed to advance this new field of research given its history of developing passive sensing devices and behavioral studies. For instance, prior studies using passive sensing find associations between passive mobile sensing and behavioral markers of mental health (137; 327; 331; 57; 333; 273; 138) and personality (223; 332), among other things. A study by Obuchi et al. (243) goes on to show the promise of passive sensing data by linking it with brain functional connectivity. The authors obtain an F1 Score of 0.79 when classifying N=105 students’ functional connectivity between two regions of the brain by just using mobile sensing data. Specific to future of work research, a work on job performance using mobile sensing (220) shows that mobile sensing can coarsely predict workplace performance. The authors collect mobile sensing and self-report data from N=554 workers in different industries across 60 days. The researchers train a machine learning model to identify if a worker is a higher or a lower performer using the self-reported labels and mobile sensing data, obtaining an AUC score of 0.83. In another work, Muralidhar et al. (226) study job performance of 100 students at a hospitality school by collecting their interview video. Participants are asked to perform a reception desk based role with “real” clients (i.e., the researchers) as a means to gauge how they would perform in the job. The researchers (226) annotate the recorded video of these interactions using five independent coders/raters. Using verbal and non-verbal features extracted, the researchers show that they are able to obtain an $R^2$ of 0.25 for inferring perceived performance and soft skills in the reception desk scenario. Also, Swain et al. (79) use passive sensing to characterize daily activities and personalities of N=603 information workers. They report that fewer location visits, batched phone-use, shorter desk-sessions and longer sleep du-
ration are related with better organizational citizenship behavior and increased task proficiency among the workers.

While there is a growing interest in assessing the workplace and job performance particularly using traditional self-reports, there has been no work to the best of our knowledge that studies changes in behavior and physiology in the workplace after promotion using mobile sensing.

Section 3.3

Methodology

In what follows, we discuss our study design, mobile sensing approach, ground-truth, demographic information of the workers in this year long study and the features used in the analysis.

3.3.1. Study Design

The Tesserae study (206) recruited N=750 information workers across different companies in the USA. All the participants are in the study for a year and respond to several surveys, as detailed in (206). Individuals participating in the study are provided with a Garmin Vivosmart 3 (190) wearable, a continuous mobile sensing app based on StudentLife (329) and a set of Bluetooth beacons to be placed at their home, on their work desk, in their office and on their keychain. The study protocol is fully approved by the Institutional Review Board (IRB). While we discuss important and relevant information from the Tesserae project as it relates to promotion, please refer to Tesserae study (206) for full details of the study design, participants, data collection, etc. Participants in the Tesserae study are instructed to maintain data compliance percentage of 80% to be eligible for monetary remuneration. In this paper, we study N=141 workers from the complete cohort of N=750 who are
promoted during the period of the study as well as the remaining 609 workers who are not promoted. Out of 141 promoted workers, 48 work at a multinational consultancy company, 44 work at a multinational technology company. The remaining 49 participants work in a software company, a university and at various other small companies.

3.3.2. Demographic Descriptors
With respect to demographic information of the promoted participants, 32% are female, 50% are male and the gender of remaining 18% of workers is unknown because they are associated with the blinded set of participants used in evaluation of the Tesserae study. 48 participants are under the age of 30, 66 are between 30 to 50 and 2 participants are above the age of 50. In terms of highest education attainment, 43% of the participants have a college degree and 33% have a postgraduate degree (Master’s or Doctorate). The remaining participants have attended some form of college or graduate school (perhaps, programs that do not award degrees). Tenure wise, the majority of the participants have been with their current employer for two years or more. Figure 3.1 shows the distribution of this demographic information.

3.3.3. Promotion Ground Truth
The ground truth for the analysis is collected through participant surveys and self-reports. At the end of the year long study, participants complete exit surveys. Two of these surveys specifically ask: Were you recently promoted? and When were you promoted? (month and year). Participants also go back through their calendar in order to submit significant life events (e.g., illness, marriage, vacation, promotion) that happened during the year long study period. Participants who choose to share the significant life events are compensated extra for doing so. The ground truth used for this paper uses responses to both the exit survey and life events survey associated
3.3 Methodology Detecting Job Promotion Using Mobile Sensing

Figure 3.1: Demographic information of the promoted participants.

with promotion. The exit survey asks for the month and year of the most recent promotion, while the life events survey asks for the exact day (from their calendar) of any significant event (in our case, promotion). As a result, we know the day of promotion for N=48 participants, while for the majority of the participants (i.e., N=93), we only know the month and the year of their promotion.

3.3.4. Mobile Sensing System

The mobile sensing app is installed on the participants iOS/Android phone and tracks participant’s phone usage, location, physical activity and Bluetooth interactions with Bluetooth beacons (at work, home, office, keychain) and runs passively without any user interaction in the background of their phone. The mobile app also syncs with the Garmin wearable streaming data from the wearable to the phone. Through the wearable we collect various physiological and behavioral data, such as, sedentary duration, motion intensity, sleep, stress, steps and heart rate data. The Garmin wearable also
connects to its own companion phone application which streams information to its web backend in order to compute daily summary information. Workplace activity and interactions related information are collected with the help of Bluetooth beacons (Gimbal). The passive sensing app implements the Gimbal’s API to capture proximity or co-location of Bluetooth beacons. This acts as a proxy in our analysis for number and duration of interactions, breaks and time spent at desk/home/work. For full details on the collection system, see (206).

3.3.5. Features

The features that we use in this paper are mostly based on prior literature discussed in the Related Work section. We list the generated features in Table 3.1. We generate daily summary features of activity (e.g., sedentary duration, distance, steps), heart rate (e.g., average heart rate, heart rate variability), stress (e.g., average stress duration), workplace behavior (e.g., number of interactions, time spent at desk) and sleep (e.g., sleep duration, Rapid Eye Movement (REM) sleep duration, duration of wake-ups). Each participant is given Bluetooth beacons to place at their home, at their work desk and on their key-chain. A Bluetooth beacon is also placed at the entrance of their respective workplace. Based on whether the signals of the respective Bluetooth beacons are detected by the participant’s phone (and for how long), we infer the time spent at work, time spent at home, time spent at their work desk, and number and duration of breaks away from their work desk. Workplace interactions are inferred based on the strength of the detected Bluetooth signals: if participant A’s phone detects a Bluetooth signal of participant B with greater signal strength (and likewise participant B’s phone detects participant A’s Bluetooth signal with greater signal strength), we count that as an interaction (if it lasts for more than a threshold of a few minutes). With respect to stress measure, it is provided by the participant’s Garmin wearable. It ranges from ‘rest stress’ with values from 0-25, ‘low stress’ with
values from 25-50, ‘medium stress’ with values from 50-75, and finally ‘high stress’ with values from 75-100. According to the Garmin specification (309), inference of the stress measure is computed by taking into account participant’s activity level, respiration rate, HR and HRV during baselines (309). Typically, when there is an increase in HR, drop in HRV lower than the resting state baseline, and respiratory rate is low relative to the HR, the Sympathetic Nervous System (SNS) which is responsible for physiological responses related to heart rate dominates, leading to a stress state (309). Sleep related data are collected via the wearable and the phone (203). In addition to the total sleep duration, the wearable provides us duration relevant to several sleep stages such as awake period, deep, light and REM sleep period. We also calculate daily sleep debt based on the difference in the ideal amount of sleep a participant gets in the prior week.

We divide the day into several “epoch” periods to better understand a worker’s physiological and behavioral sensing data over different periods of the day for modelling purposes: epoch 0 (representative of entire 24 hour day), epoch 1 (12 am - 9am; night time, typically when people sleep), epoch 2 (9am - 6pm; day time, typically when people work) and epoch 3 (6pm - 12am; evening time, typically when people leave work and go home or elsewhere). We also consider other time periods based on Bluetooth beacons, such as, time spent at their desk (while at work), time spent not at their desk (while at work), and time spent while not at work (i.e., before they arrive and after they leave). Features related to activity, stress, sleep, steps and heart rate allow us to analyze the effect of promotion on wellbeing, while workplace behavior, distance travelled and phone usage based features are important from the perspective of capturing changes in employee engagement, communication and socialization. Prior work has shown that promotion can impact all these aspects of wellbeing (147; 83; 166; 46), therefore, we use all of these features in our modelling.
3.4 Analysis

We discuss our modeling and analysis of physiological and behavioral patterns associated with promotion periods in what follows. We present our results in the next section.

Table 3.1: Features generated from the mobile sensing system: from the phone, wearable and beacons.

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Sedentary duration, physical activity, motion intensity</td>
</tr>
<tr>
<td>Workplace activity</td>
<td>Time spent at desk, time spent at work, number of times participants leave their desk (for durations of 5, 15 and 30 minutes), arrival time to work, departure time from work, duration spent on breaks</td>
</tr>
<tr>
<td>Workplace interactions</td>
<td>Number of unique participants, duration spent on interactions, percent of time spent alone, percent of time spent in different interactions (interacting with only one person, with one or more person, with two or more person, or three or more person)</td>
</tr>
<tr>
<td>Distance</td>
<td>Total distance travelled, average distance from home, number of locations visited</td>
</tr>
<tr>
<td>Stress</td>
<td>Duration of high/medium/low stress, average stress levels</td>
</tr>
<tr>
<td>Sleep</td>
<td>Sleep duration, daily sleep debt, duration of deep/light/REM sleep, duration of wake-ups</td>
</tr>
<tr>
<td>Steps</td>
<td>Number of steps, walking/running durations</td>
</tr>
<tr>
<td>Heart rate</td>
<td>Average heart rate, heart rate variability</td>
</tr>
<tr>
<td>Phone usage</td>
<td>Number of unlocks, usage duration</td>
</tr>
</tbody>
</table>
3.4 Analysis

3.4.1. Analyzing Changes in Behavior and Physiology

We start our analysis by exploring the changes in behavior and physiology of the participants after promotion. We do this by performing a paired Wilcoxon signed rank test (340) on the sensing data of the participant up to 60 days before and 60 days after the month of promotion. We use Wilcoxon signed rank test because the data points are non-normally distributed. Wilcoxon signed rank test compares two related samples to detect whether there's a difference between them after an intervention (in our case, the intervention is promotion.)

As mentioned in the Introduction section we select a 60 day epoch period for analysis in order to maximize the number of participants we can include in the analysis across the year based on the amount of data we have from all the promoted workers. We know the day of promotion for N=48 participants, while for the majority of the participants (i.e., N=93), we only know the month and the year of their promotion. Our analysis is based on N=141 of these participants who reported being promoted during the period of the study. Because of this, we exclude the month of the promotion from the analysis. Let us explain that last sentence. If a participant reported being promoted in February, for example, we exclude all their data for February from our analysis irrespective of which specific day they are promoted on. We do this because we do not know the exact day of promotion for a majority of the participants (i.e., N=93) and so the only way we can avoid data before promotion (i.e., non-promoted period) "leaking" into data after promotion (i.e., promoted period) is by removing data associated with the promotion month which we know for all N=141 of these participants. This means that non-promoted period includes data just before the month of promotion (i.e., until the end of January in our example of promoted in February) and promoted period includes the data after the month of promotion (i.e. from the start of March ).

We report result of the analysis using four criteria: 1) the overall behavioral and physiological change
among all the promoted workers; 2) the difference in behavior and physiology between promoted workers after promotion and non-promoted workers; 3) the behavioral and physiological changes associated with gender; and finally 4) the changes associated with high and low performers (220).

3.4.2. Predictive Classification Model

For the final part of the analysis, we build a predictive classification model that utilizes the sensing data from workers’ phone, wearable and beacons (associated features discussed in Table 3.1) to determine whether the data belongs to promoted or non-promoted periods. Our aim is to be able to detect promotion events by leveraging machine learning technique. We include both the promoted workers and non-promoted workers in this analysis.

For promoted workers, we label the data after the month of their promotion as belonging to promoted class, and data before the month of their promotion as belonging to non-promoted class. Similarly, in case of non-promoted workers, we label all their data as belonging to non-promoted class. Then, we design a model that can detect which period the sensing data belongs to. For the purposes of modeling, we treat the problem as a multivariate time series classification task. It could be the case that multiple sensing streams (e.g., arrival time to work, number of steps, sleep duration) interplay with each other and if we can capture these variations, we would be equipped with better cues to detect promotion based on the sensing data. For each stream, we create a daily summary time series spread over 7 days (i.e., a week) for each of the promoted and non-promoted periods. We make predictions on each weekly grouping of data and once we have the entire month’s prediction (i.e., 4 weeks prediction), we use majority voting technique to identify whether that month belongs to a promoted or a non-promoted period. Our machine learning model makes use of a recent technique in time series classification which the authors report to have
outperformed other state of the art time series classification algorithms (86).

The approach used by ROCKET (Random Convolutional Kernel Transform) (86) is influenced by the success of Convolutional Neural Network (CNN) (178). Although CNNs are mostly known for their performance in image based tasks (116; 299), they also perform well when used in time series classification (97; 96). Researchers believe that they could be effective in time series classification because the convolutional kernels capture patterns in the input time series (96). A kernel refers to a matrix that is convoluted with the input time series through a sliding dot product (96). The result is called feature map, another matrix, which is used for classification. Kernels can capture complex patterns and shapes reflected in the feature map. Feature maps are the result of applying the kernels to an input. In essence feature maps are a rendition of the input showing how much of the pattern represented by the kernel are present in the input time series. However, unlike CNNs where the convolution kernels are typically learned, ROCKET uses random convolutions. What this means is that, the kernel weights are not learned, they are randomly generated.

Random convolutions are not novel innovations; they have been used in prior works (284; 144; 258; 146). In fact, researchers have suggested that random convolutions may be advantageous for small datasets where learning better kernel representation is difficult (144; 354). ROCKET performs random convolutions over the time series data to extract patterns which are then used by a linear model to make predictions. This approach provides significant gains over the state of the art performance with huge reductions in time and computation power required (86). ROCKET kernels have random length, dilation, padding, weights and bias (121). The only hyperparameter for ROCKET is the number of random kernels to generate. For more detail on ROCKET, please refer (86). In our implementation of ROCKET we extend the model to take into account the multivariate nature of our dataset. We do this by performing
3.5 Result

Detecting Job Promotion Using Mobile Sensing

a multi-channel 1 dimensional convolution over the time series data. Multi-channel here refers to the number of sensing streams. We perform a one dimensional random convolution over all the sensing streams grouped by 7 days. Thereafter, we select features using Sequential Forward Selection (SFS) (102) approach for classification.

Section 3.5

Result

In what follows, we present our results with regards to behavioral and physiological change in workers after promotion (i.e., promoted period). We report results associated with specific subgroups of interest including gender and high/low job performers. We also discuss our results on detecting job promotion using mobile sensing.

3.5.1. Behavioral and Physiological Change after Promotion

Within Promoted Workers. We provide initial insights into behavioral and physiological differences captured through mobile sensing by comparing features from up to 60 days period after the month of promotion and up to 60 days before promotion. We use paired Wilcoxon signed rank test for this purpose because it allows us to compare paired data for each promoted worker before and after promotion. We show the significant behavioral and physiological changes in Table 3.2. Note, all the reported changes are significant with p-value of less than 0.10 after correcting for multiple comparisons using false discovery rate using the Benjamini Hochberg (FDR-BH) procedure (35).

As shown in Table 3.2, we find that the workers are less physically active after being promoted. Given the fact that most of our cohort comprises information workers requiring knowledge work (i.e., problem solving, cognitive skills) rather than physical work, they likely spend longer periods of time sitting and working which results in
Table 3.2: Changes in behavior and physiology of workers after promotion. The following table lists significant changes in behavior and physiology we find for all the participants after promotion. Epochs refer to grouping of different periods of the day; epoch 0: 24 hours (whole day); epoch 1: 12 am - 9 am (night/early morning); epoch 2: 9 am - 6 pm (working hours); epoch 3: 6 pm - 12 am (evening). Note, all the reported changes are significant with p-value of less than 0.10 after correcting for multiple comparisons using false discovery rate using the Benjamini Hochberg (FDR-BH) procedure (35).

<table>
<thead>
<tr>
<th>Category</th>
<th>Period</th>
<th>Change in pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Epoch 0</td>
<td>Total physically active duration reduces after promotion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time spent in vehicle decreases</td>
</tr>
<tr>
<td></td>
<td>Epoch 2</td>
<td>Time spent in vehicle decreases</td>
</tr>
<tr>
<td>Workplace activity</td>
<td>Epoch 0</td>
<td>Increase in number of 15 minutes break</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decrease in number of 30 minutes break</td>
</tr>
<tr>
<td>Workplace interactions</td>
<td>Epoch 0 / 2 / 3</td>
<td>Number of interactions decreases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Duration of interactions decreases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of unique participants in conversation decreases</td>
</tr>
<tr>
<td>Distance</td>
<td>Epoch 2</td>
<td>Fewer number of unique locations visited</td>
</tr>
<tr>
<td>Stress</td>
<td>Epoch 0-3</td>
<td>Decrease in high and low stress duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Increase in rest stress duration</td>
</tr>
<tr>
<td></td>
<td>Epoch 0 / 2 / 3</td>
<td>Increase in stress variability</td>
</tr>
<tr>
<td>Sleep</td>
<td>At night</td>
<td>Increase in sleep duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Increase in REM sleep duration</td>
</tr>
<tr>
<td>Steps</td>
<td>Epoch 0 / 3</td>
<td>Decrease in step count</td>
</tr>
<tr>
<td>Heart rate</td>
<td>Epoch 1</td>
<td>Increase in heart rate</td>
</tr>
<tr>
<td></td>
<td>Epoch 2 / 3</td>
<td>Decrease in heart rate</td>
</tr>
<tr>
<td></td>
<td>At desk</td>
<td>Decrease in HRV while at desk</td>
</tr>
<tr>
<td>Phone usage</td>
<td>Epoch 1</td>
<td>More time spent on phone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Increase in the number of unlocks</td>
</tr>
<tr>
<td></td>
<td>Epoch 3</td>
<td>Increase in the number of unlocks</td>
</tr>
</tbody>
</table>

Wilcoxon paired test; p-value < 0.10 (FDR-BH)

less physical activity. This is corroborated in the table by a reduction in step count as promoted workers move through the ranks. On the other hand, we find that wellbeing related factors such as stress and sleep improve after promotion. Findings from prior research on promotion and wellbeing are inconclusive; some researchers show that
promotion can lead to better health outcomes, while others find promotion leads to deteriorating health outcomes. In a panel data of $N=2,681$ workers, Johnston et al. (147) report finding small yet positive association with job stress and promotion (i.e., as people get promoted, their job stress also increases). The authors (147) do not find any significant effect of promotion on general health and happiness. Boyce et al. (46) find that people suffer significant deterioration in their psychological wellbeing after promotion. However, Karasek et al. (150) argue that higher occupational levels that are achieved as a result of promotion lead to reduced stress because employees have more autonomy and control of their work which helps them mitigate the high demands of their job. In another line of research, job satisfaction is shown to be strongly associated with employee wellbeing (293), and because promotion typically leads to an increase in job satisfaction (106; 278), it could be inferred that promotion leads to better wellbeing through the mediating role that it plays in increasing job satisfaction.

Mobile sensing (HRV from the wearable) and location information (from the beacons) allow us to measure deeper contextual information associated with workers not readily available to prior researchers; that is, we can analyse a worker's behavioral and physiological signals while they sit at their desk, for example. We find that although the stress reduces for promoted workers across different periods of the day, their heart rate variability (HRV) decreases while they are at their desk. Heart rate variability is a commonly accepted biomarker of stress and generally a lower HRV is associated with higher stress (156). We believe that the decrease in HRV while at the work desk is potentially pointing to the fact that their stress is increased while they are working at their desk maybe because of the added responsibility of their new role after promotion. However, when considering different periods of the day (i.e., epochs) as well as the entire day, we find that promotion actually leads to positive health outcomes.
in terms of stress and sleep. Perhaps this is because the job rewards (job flexibility, control, autonomy, salary etc.) actually help people cope with the increased demands and responsibility and maximize their overall health when considered in relation with the entirety of the day or other periods other than while working at their desk.

In terms of workplace activity, we find that there is an increase in the number of breaks of shorter duration and a decrease in breaks of longer duration. It is not clear why this would be the case. One possible explanation is that the promoted workers are highly engaged at work as a result of changing demands in their new roles. Prior research shows that workplace promotion typically leads to an increase in employee performance and productivity (262; 278). This might relate to another one of our findings – people visit fewer places during the working hours once promoted in comparison to the period before promotion. We posit that workers might be more engaged with new work activities as a result of promotion, which is discussed in the literature (83).

Workplace interaction with colleagues appears to go down after promotion, in both the numbers of interactions as well as the duration of interactions. This is an interesting finding. In addition, after promotion people tend to interact with fewer people. There are multiple interpretations of this result. For one, we observe that promoted employees reduce the number of longer breaks, which might imply that they are engaging more in work items and as a result reducing time spent interacting with other workers. Another interpretation of this result is that promoted workers simply experience changes in communication and engagement with former peers (e.g., individuals, group members, supervisors, managers). The literature indicates that non-promoted workers (for instance, after getting passed over for promotion) may be less attached to their co-workers, hence decreasing the overall interaction for promoted employees (118). In a survey of N=20 promoted employees, Kramer et al. (166)
3.5 Result

Detecting Job Promotion Using Mobile Sensing

Figure 3.2: Comparison of changes after promotion. The figure shows a number of the sensing streams along with the percentage change in them after promotion. Comparisons are shown between median values of the streams. The green dot represents the sensor reading before promotion and red dot, after promotion. We can see that some streams such as average stress level decrease, while others such as HRV increase. All the duration are represented in minutes.

Figure 3.2: Comparison of changes after promotion. The figure shows a number of the sensing streams along with the percentage change in them after promotion. Comparisons are shown between median values of the streams. The green dot represents the sensor reading before promotion and red dot, after promotion. We can see that some streams such as average stress level decrease, while others such as HRV increase. All the duration are represented in minutes.

find that employees experience a sense of social isolation from peers and supervisors after getting promoted. After promotion, employees report feeling isolated because "they were no longer one of them". It makes sense that communication channels, conversational engagement and other workplace interaction are disrupted once an individual is promoted.

Finally, we find that after promotion, workers spend more time on their phone during the night/early morning period. We also capture more number of unlocks during night/early morning and evening periods, both of which are periods after work. Figure 3.2 shows a number of the significant features that change along with the percentage change after promotion. Comparisons are shown between median values of the streams in the figure. The green dot represents the sensor reading before promotion and red dot after promotion. We can observe that some streams such as average stress level decrease, while other streams such as HRV increase. Our results
find significant common patterns across the cohorts we study. While the literature supports a number of our observations we do not claim that these observed behavioral and physiological changes are generalizable.

**Between Promoted and Non-Promoted Workers.** From the previous analysis, we learn that promoted workers’ physiology and behavior changes after promotion. However, we only took within-person changes into account. In what follows, we study the significant differences in behavior that show up when we compare promoted participants with non-promoted participants (i.e., studying between-person differences).

Similar to our prior analysis, we observe up to 60 days of data after promotion period for promoted workers, and similarly 60 days of data for non-promoted workers. We then perform a between-groups comparison using Wilcoxon rank-sum test (also known as Mann-Whitney U test) (198). The Wilcoxon rank-sum test is a non-parametric version of independent T-test, allowing us to compare two independent groups. We find several significant differences between the promoted and non-promoted workers, as listed in Table 3.3.

We find that promoted workers are less physically active than non-promoted workers. They have a comparatively lower physically active duration and take fewer steps throughout the day. Non-promoted workers seem to travel more and also venture out further away from their home in comparison to promoted workers. Furthermore, promoted workers spend more time by themselves and at their work desk. This supports our finding from the within-person analysis that we discussed before. As mentioned in Section 3.5.1, one reason behind being less physically active could be straightforward: the vast majority of our study participants are information workers, therefore, a rise through the rank might coincide with spending more time stationary at their desk or in meetings which result in less physical activity. We also find that promoted workers sleep longer and have lower average stress throughout the day. Not surprisingly (since
they have lower average stress), they have a higher HRV as well. Most of the results we obtain from this between-groups comparison corroborates our findings from the earlier section. We learn that there are significant differences not just within the behavior of the promoted workers but also between the promoted and non-promoted participants. In Section 3.5.4, we leverage our findings from these two analyses to train a machine learning model that can learn patterns from these differences.

Table 3.3: Differences in patterns of non-promoted/promoted workers. The table below lists the differences in behavior and physiology of non-promoted/promoted workers (after they get promoted). There are 141 promoted workers and 609 non-promoted workers in our analysis.

<table>
<thead>
<tr>
<th>Group</th>
<th>Differences in patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoted</td>
<td>Time spent in vehicle is lower during overall day.</td>
</tr>
<tr>
<td></td>
<td>Physically active duration is lower throughout the entire day.</td>
</tr>
<tr>
<td></td>
<td>More time spent alone and at work desk while in the workplace.</td>
</tr>
<tr>
<td></td>
<td>Higher rest stress duration and lower average stress.</td>
</tr>
<tr>
<td></td>
<td>Higher sleep duration.</td>
</tr>
<tr>
<td></td>
<td>Average heart rate variability while at desk and also while not at work is higher.</td>
</tr>
<tr>
<td></td>
<td>Total number of phone unlocks in the overall day is higher.</td>
</tr>
<tr>
<td>Non-promoted</td>
<td>More distance travelled throughout the day.</td>
</tr>
<tr>
<td></td>
<td>Travel further away from home than promoted participants in the entire day.</td>
</tr>
<tr>
<td></td>
<td>Lower REM sleep duration.</td>
</tr>
<tr>
<td></td>
<td>More number of steps throughout the day.</td>
</tr>
<tr>
<td></td>
<td>Average heart rate at work is lower.</td>
</tr>
</tbody>
</table>

Wilcoxon rank-sum test; p-value < 0.10 (FDR-BH)

3.5.2. Subgroup Analysis: Gender and Job Performance

Analysis of promotion on the entire cohort might mask important effects that occur for certain subgroup of participants. In order to investigate the differences within subgroups, we perform further analysis of behavior and physiology change considering gender and job performance.
3.5 Result

**Detecting Job Promotion Using Mobile Sensing**

**Gender.** Out of 141 participants, there are 45 females and 71 males who are promoted. The remaining 25 participants are blinded (that is, we do not know their gender). Table 3.4 lists the changes we find in participants based on gender.

Table 3.4: Changes in patterns after promotion when considering gender. We observe the change in behavior and physiology of promoted workers based on gender. There are 71 males and 45 females who are included in our analysis. Note, all the reported changes are significant with p-value of less than 0.10 after correcting for multiple comparisons using false discovery rate using the Benjamini Hochberg (FDR-BH) procedure (35).

<table>
<thead>
<tr>
<th>Gender</th>
<th>Changes in patterns after promotion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stress is more variable during the entire day, in the working hours (9am - 6pm) and during the evening (6pm - 12am). Duration with low stress decreases during all periods (entire day, morning, working hours and evening). Medium stress duration increases during working hours (9am - 6pm). Average workplace interaction duration decreases across the entire day as well as during the evening (6pm - 12am). Total distance travelled increases for the entire day. Heart rate increases during evening (6pm - 12am). Heart rate variability decreases while not at work.</td>
</tr>
<tr>
<td>Female</td>
<td>Phone usage during night/early morning (12 am - 9 am) increases. Number of unique locations visited during working hours (9am - 6pm) decreases. Number of unlocks in the entire day, during night/early morning (12 am - 9 am) and during evening (6pm - 12am) increases.</td>
</tr>
<tr>
<td>Male</td>
<td>Average stress level in the entire day, during night/early morning (12 am - 9 am) and during evening (6pm - 12am) decreases. Total step count in the evening (6pm - 12am) decreases. REM sleep duration increases. Number of workplace interactions and unique participant decreases throughout the day (overall day, working hours and evening). Variation of heart rate while at the work desk increases.</td>
</tr>
</tbody>
</table>

Wilcoxon paired test; p-value < 0.10 (FDR-BH)

Our gender based analysis shown in Table 3.4 clearly indicates that there are significant differences between females and males in terms of behavior and physiology changes after promotion. This is an interesting result. We find that male workers show an increase in phone usage during night/early morning whereas we do not see similar changes for promoted female workers. Male workers show a decrease in the
number of unique locations visited during working hours. For female workers, we find that the total distance travelled increases when we consider the entire day. The most significant changes for female workers are related to stress and heart rate variability. While we find that the average stress level decreases for male workers, females experience an increase in medium stress duration during working hours. Nyberg et al. (242) report that female workers’ self-reported health deteriorates at a higher rate in comparison to male workers after getting promoted. Prior studies report that female workers with same position in a company as male workers have higher demands (321), less financially rewarding jobs (38; 241), larger sacrifices in private life (240) and lack of social rewards (133) which ultimately leads them to suffer more stress in jobs with higher occupational status in comparison to males. We also find that the female workers’ HRV while not at work decreases. We do not find this for males. Previous work shows that female workers are more exposed to work-family conflicts and ‘double burden’ (241; 321; 93), which might be the reason why they experience a decrease in HRV while they are not at work. Interestingly, the number of workplace interaction seems to decrease for both genders, possibly reflecting that both males and females experience increased engagement in the workplace, or, workers experience a change in communication, regardless of gender.

**Job Performance.** In this subgroup analysis, we first make an attempt to uncover the association (if any) in our dataset with respect to job performance and promotion. We then analyze the behavioral and physiological changes.

As discussed in Section 3.3, the Tesserae study (206) collects self-reports along with mobile sensing data from N = 750 participants. One of the many self-reports collected from the participants includes job performance metrics every 3 days for the first 60 days of the year long study. We want to see if the self-reported performance metric is an indicator of future promotion. The participants self-report their performance
using four metrics/surveys: counterproductive work behavior (CWB) (270; 272), organizational citizenship behavior (OCB) (71; 270), in-role behavior (IRB) (44; 342) and individual task performance (ITP) (44; 123). For our analysis, we take an average of the participant’s self-reported performance scores for these four metrics.

We cluster the participants into higher and lower performers based on the four self-reported performance metrics proposed in (220). In this approach (220), multiple iterations of K-means clustering is run with initial centroid set to the maximum values of OCB, IRB and ITP survey scores whereas the minimum value of CWB survey score is used. The idea here being, OCB, IRB and ITP are positive performance metrics, whereas CWB is an indicator of negative workplace behavior (i.e., the higher the aggregated value of CWB response, the lower the job performance). Hence, K-means clustering with K=2 and initial centroid set to maximum value of positive indicators of job performance and minimum value of negative indicators of job performance should ideally lead to a division of participants into two classes: “high performers” (those that have high OCB, IRB and ITP but low CWB) and “low performers” (high CWB, but low OCB, IRB and ITP). After we cluster the workers into these two classes (viz. high performer and low performer), we investigate whether there is any correlation between the two performance groups and promotion. The question that drives this exploration is “Is prior job performance an indicator of future promotion?”.

We find that the Mathew’s Correlation Coefficient (MCC) (205) on performance class (higher/lower performers) and being promoted is -0.06. This indicates that there is no relationship between performance class and being promoted. For what it’s worth, 72 of our 141 participants fall in high performer’s class while 44 are in the low performer’s class. The remaining 25 participants are blinded; we do not know their performance metrics. In order to investigate the influence of individual performance metrics on promotion (as opposed to the broad higher/lower performance class based
on clustering), we fit a logit model with the performance metrics and performance class as independent variables and being promoted (or not) as dependent variable. However, again we do not find any statistically significant relationship. All this leads us to conclude that self-reported performance is not reflective of future promotion in our dataset.

Table 3.5: Changes in patterns of higher/lower job performers. The table below lists the changes in behavior and physiology of higher/lower job performers after they get promoted. There are 72 higher performers and 44 low performers in our analysis.

<table>
<thead>
<tr>
<th>Performance Class</th>
<th>Changes in patterns after promotion</th>
</tr>
</thead>
</table>
| **High Performers** | Number of unique locations visited during working hours (9am - 6pm) and overall day decreases.  
Stress has more variability during overall day, working hours (9am - 6pm) and during evening (6pm - 12am).  
Average interaction duration in evening (6pm - 12am) decreases.  
Average heart rate in overall day, during working hours (9am - 6pm) and during evening (6pm - 12 am) decreases.  
Average HRV increases in overall day.  
Number of unique participants in interactions with during the overall day increases, while it decreases during working hours (9am - 6pm) and evening (6pm - 12am). |
| **Low Performers** | Stress has more variability during working hours (9am - 6pm) and during evening (6pm - 12am).  
Stress duration decreases in overall.  
Number of unique participants goes down in the entire day, working hours (9am - 6pm) and evening (6pm - 12am).  
Heart rate increases during night/early morning (12am - 9am). |

Wilcoxon paired test; p-value < 0.10 (FDR-BH)

Finally, we investigate the within-person changes of high performers and low performers after getting promoted. Table 3.5 shows the significant changes we observe. Perhaps what is more surprising about our findings is that higher performers have an increase in number of unique participants that they interact with in the overall day. These higher performers show an improvement in their HRV and heart rate. Stress reduces for both the groups after promotion but there is an increase in stress variability. For low performers, we find that heart rate increases during night/early morning, after leaving work for the day.
3.5 Result Detecting Job Promotion Using Mobile Sensing

(a) Light sleep duration changepoint

(b) Number of 15 minute breaks changepoint

(c) Stress duration while at work changepoint

Figure 3.3: Changepoint detection for several sensing streams (viz. sleep, breaks, stress). The figure shows changepoint detected for three different promoted workers for ‘light sleep duration’, ‘number of 15 minute breaks’ and ‘stress duration at work’ streams during one of the weeks in the month of their promotion. The vertical dashed line denotes the changepoint identified – the blue and red shaded regions represent the period before and after the changepoint, respectively. The sensing time series figures clearly show that there is a significant variation in the light sleep duration (i.e., decrease), number of 15 minute breaks (i.e., increase) and stress duration at work (i.e., increase) following the changepoint. Note, the changepoint time series selected are represented on many other examples and are just shown here to illustrate the visible change in physiological and behavioral patterns.

3.5.3. Changes in Individual Streams

We extract numerous features from our sensing dataset sourced from workers’ phones, wearables and Bluetooth beacons. Before we discuss our predictive classification model, we want to explore if individual sensing streams change around the time of promotion. We use offline changepoint detection to study this. Changepoint detec-
3.5 Result

Detecting Job Promotion Using Mobile Sensing

tion (10) is an approach used to identify whether there is a change in a sequence of observations over time. Aptly named, a changepoint is a point at which there is an abrupt variation in the time series data. In offline changepoint detection (318), changes are identified by taking into account the entire dataset in a retrospective manner. Online changepoint detection (10) is used to detect changes in real time settings. Because we are exploring the streams after the collection of the dataset, we use offline changepoint detection to look back in time to recognize where the change happened. Note, that this is a search based technique; it is not using machine learning.

We first aggregate all the available sensing streams by week to create a weekly time series for each stream and then we pass each of the time series streams through binary segmentation search method (112). Binary segmentation based search method works by trying to identify a single changepoint in the entire dataset in the first instance and then by breaking the time series up into further smaller splits to come up with the most significant changepoints. We use weekly summarised time series of each sensing streams for up to 20 weeks including the month of the promotion. We then investigate the composition of streams that change the most during the month of the promotion (i.e., the streams that have changepoints during that period). In Figure 3.3, we show three examples of changepoint detected on three separate streams (viz. light sleep duration, number of 15 minute breaks and stress duration at work). We observe that there is a significant change in each of the streams following the week where changepoint is identified. The vertical dashed line in the time series figures denotes the changepoint – the blue and red shaded regions represent the period before and after the changepoint, respectively. These sensing time series figures clearly show that there is a significant variation in the light sleep duration (i.e., decrease), number of 15 minute breaks (i.e., increase) and stress duration at work (i.e., increase) following the changepoint. Note, the changepoint time series selected are represented on many
other examples and are just shown here to illustrate the visible change in physiological and behavioral patterns. For the entire cohort, we report the result by categorizing all the sensing streams (i.e., feature categories) into 5 groups: heart rate and heart rate variability (HRV), stress, sleep, workplace behavior (includes workplace activity and interactions) and activity (includes everything else: distance, activity, phone usage).

![Figure 3.4: Composition of different streams. The figure shows the composition of different groupings of streams towards changepoint detection. Each blue colored radial in the image represents a participant in our study. The closer the radials are towards a particular group, the more changepoints were detected by that group of streams for that particular person. We find that in average, 22% of changepoints are detected on Stress related streams, 21% on Activity related streams, followed by 20% on Sleep, 20% on HR & HRV and 17% on Workplace behavior related streams.](image)

In Figure 3.4, we show which of these groups change the most during promotion. As can be observed in the figure, stress based streams change the most (i.e., on average, 22% of the changepoints are identified through stress based streams). Activity (21%), sleep (20%) and HR/HRV (20%) based streams also change during the period of promotion. Streams related to workplace interactions and desk activity (17%) do not have as many changepoints in comparison to these more significant streams. In Table 3.6, we list a number of the individual sensing streams within each groups that
change the most during promotion.

Table 3.6: Streams that change the most. In this table, we list some of the individual streams that changed the most during the period of promotion.

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Total step count</td>
</tr>
<tr>
<td></td>
<td>Calories burned</td>
</tr>
<tr>
<td></td>
<td>Highly active duration</td>
</tr>
<tr>
<td>Workplace behavior</td>
<td>Time spent at work</td>
</tr>
<tr>
<td></td>
<td>Average break duration</td>
</tr>
<tr>
<td></td>
<td>Departure time from work</td>
</tr>
<tr>
<td>Stress</td>
<td>Duration of high stress</td>
</tr>
<tr>
<td></td>
<td>Duration of medium stress</td>
</tr>
<tr>
<td></td>
<td>Duration of low stress</td>
</tr>
<tr>
<td>Sleep</td>
<td>REM sleep duration</td>
</tr>
<tr>
<td></td>
<td>Daily sleep debt</td>
</tr>
<tr>
<td></td>
<td>Total sleep duration</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>Average HR at desk</td>
</tr>
<tr>
<td></td>
<td>Average HR not at work</td>
</tr>
<tr>
<td></td>
<td>Average HRV at desk</td>
</tr>
</tbody>
</table>

3.5.4. Predictive Classification Model

A key goal of our study is to train a machine learning model such that it can detect the promotion period (whether promoted or non-promoted) that the sensing data is associated with. The exploratory analysis that we performed in the earlier section hinted towards the fact that passive sensing streams change during or after the period of promotion. With a predictive machine learning model, we investigate if using all these changes and variations of multiple streams together can give us a good performance. As mentioned in Section 3.4, we treat the problem as a time series classification task. Basically, for each promoted and non-promoted period, we first group the data into a 7 days time series. We extract features from the 7 days time series and feed it into the model. Once we have the predictions for 4 weeks, we perform a majority voting over all the 4 individual predictions to say whether the
corresponding month belongs to a promoted period or a non-promoted period. In case of a tie (i.e., 2 weeks being predicted as non-promoted period and 2 weeks as promoted period), we break it by predicting class 0 (i.e., non-promotion), which is to say that we always require a majority in order to detect promotion. As we may not be able to detect promotions if we use a time series generated over a larger period of time because the important patterns may then be masked by having data aggregated over a longer period, we decided on using weekly time series instead of monthly and do majority voting over them for a month during testing. Weekly time series provides us a better granularity than a monthly time series allowing us to extract patterns that provide a more “cleaner” picture of promotion related events.

The process begins by first extracting features using 250 random kernels. The length of the kernels we use is randomly selected from \{3, 5, 7\} with equal probability. Other parameters are initialized as mentioned by Dempster et al. (86); weights are sampled from normal distribution, bias is sampled from a uniform distribution \(b \sim U(-1, 1)\), dilation is sampled on an exponential scale \(d = \lfloor 2^x \rfloor, x \sim U(0, A)\) where \(A = \log_2 \frac{l_{\text{input}} - 1}{l_{\text{kernel}} - 1}\) and padding is applied randomly with equal probability. Stride is always set to one. We implement ROCKET with the help of the PyTorch library (253) by performing a 1D convolution, where the number of channels is equal to the number of features, and the sequence length for the time series is equal to 7 (since we are extracting features for a week at a time). From the resultant feature maps, ROCKET (86) then computes two aggregated features – the maximum value (which is equivalent to global max pooling (248)) and the proportion of positive values. Overall, there are 500 features at the end. Let us provide an example of the process. For instance, we start with a 3D tensor, say, of size \((432, 133, 7)\). Here, 7 refers to the number of columns representative of each day of week. Since we are predicting for a week at a time, we create a time series of 7 days so that we can
extract features based on each week’s data. The row size in the tensor refers to the number of features which we suppose to be 133 in this case. Let us also assume that we have 432 number of weeks available in total of all the participants. Thereafter, we generate 250 random kernels of size 3, 5 or 7 where each have equal probability of occurrence. The weights, bias, padding and dilation are also randomly generated based on the settings mentioned earlier. Once we generate 250 kernels, we then perform 1D convolution over all these 432 different arrays of size (133, 7), using the randomly generated kernels. After performing convolution operations between each kernel and the weekly arrays, we obtain feature maps. From each feature map, we generate two aggregated features – the maximum value and the proportion of positive values. So basically, after performing 1D convolution using every kernel, we extract two features per kernel. At the end we have an array of size (432, 500), where 432 is the number of rows representative of each week of the participants and 500 is the number of columns representative of the features that we generated from the feature map (since we started with 250 kernels, we have 500 features each for every week – twice the number of kernels.) From an input tensor of size (432, 133, 7), we end up with an array of size (432, 500) after feature extraction. We then perform feature selection over these 500 features.

As suggested in (86), we normalize the data per sample before performing convolution operations. Participants were instructed to have 80 percent compliancy in order to qualify for monetary remuneration. However we still have numerous missing values in the dataset – some participants with more and some with less (202). We perform imputation in order to handle these missing values; we use backward filling and forward filling based on the day of the week and iterative imputation based on Random Forest (also known as MissForest (303)). Then Sequential Forward Selection (SFS) approach is used to select features. Figure 3.5 shows the performance of
3.5 Result

Detecting Job Promotion Using Mobile Sensing

Figure 3.5: Performance of models based on feature selection. The figure shows comparison of 5 different Machine Learning models with sequential forward selection. The x-axis is the number of features and the y-axis is the average AUC obtained as a result of leave subjects out cross validation. We can see that all the algorithms show an improvement in their performance as the number of features increase. The figure above shows a particular snapshot of feature selection when number of features is equal to 100. Eventually the models reach saturation point and there is a gradual drop in performance. The highest performance is obtained by Logistic regression.

different models as SFS proceeds with feature selection. Note that the figure shows a snapshot of one of the iterations limited to only first 100 features. Among the five different models that we try (viz. logistic regression, support vector machine, gaussian naive bayes, random forest and k nearest neighbor), we obtain the best performance with logistic regression. Logistic regression leads to an average AUC of 0.72 with 100 features. In Figure 3.6, we show a comparison of the performance of these different models when used with and without features generated by ROCKET.

To avoid overfitting and data leakage issues, we perform Leave 10 Subjects Out Cross Validation, which means that we do not use the same subject’s records split in train and test set at the same time. Each participant has several months of data (as mentioned earlier, for promoted workers, we do not use the month of the promotion in the analysis), some of them belong to class 0 (referring to non-promoted period) and the remaining belong to class 1 (referring to promoted period). In case of the
3.5 Result  Detecting Job Promotion Using Mobile Sensing

Figure 3.6: With and without the use of ROCKET. The figure shows the performance of machine learning models when they’re used with and without the features generated by ROCKET. "Without ROCKET features" refers to the original features. We see that features generated by ROCKET improve the performance compared to original features except in the case of K Nearest Neighbor classifier. Logistic regression outperforms other models in terms of performance.

non-promoted workers, all their data points will be labelled as 0, while the promoted workers will have a mix of class 0 and class 1 labels, depending on when they were promoted. As we have a larger number (609) of non-promoted workers than we have the promoted ones (141), we divide the non-promoted workers into folds each consisting of 100 non-promoted participants. We then repeat the predictive analysis for each fold. Note that we have 241 participants in each iteration of analysis – 100 varying non-promoted participants and 141 fixed promoted ones. This division allows us to handle two things: firstly, it allows us to deal with data imbalance issues because with this division we have similar number of promoted and non-promoted workers as opposed to having large majority of non-promoted workers (i.e., 609) if we consider all of them. Secondly, using different groupings of participants makes the analysis more robust as we are not just relying on a certain group of participants. In addition, we perform another set of predictions to further validate the generalizability of our model. We divide the dataset into stratified K folds where K = 75. Here each fold has
10 participants consisting of a combination of promoted and non-promoted workers. As a result of using stratified K fold, the distribution of promoted and non-promoted workers in each fold is similar to that of the entire dataset. At every iteration, we train on 74 folds and perform testing on the remaining fold. In this manner, we evaluate our model in such situations where we may have a large number of non-promoted workers but a fewer number of promoted workers. In essence, we make predictions on an imbalanced dataset because the held-out stratified fold has similar class distribution as that of the entire dataset which is heavily imbalanced. The result is Leave 10 Subjects Out Stratified Cross Validation. With Logistic Regression, we obtain an average AUC of 0.68 with this procedure, which is not too far off of the result we obtain from our earlier analysis (i.e., AUC of 0.72).

We use ROCKET generated features in our predictive models. As discussed previously, ROCKET has only one hyperparameter: the number of random kernels. The number of random kernels essentially defines the number of features extracted from the given signal. ROCKET extracts two features per feature map. Therefore, there will be twice the number of features as there are the number of random kernels. We extract 500 features using 250 kernels. In figure 3.7, we show the distribution of learned weights when we do not use SFS for feature selection and also the distribution when we select top 100 features based on SFS. We can clearly see that SFS leads to coefficients that are non-zero. The coefficients, as shown in the figure, are obtained when we train a Logistic Regression model based on features selected by SFS. This shows that feature selection is promising and it helps isolate the key features that can improve the performance of the model.
3.6 Discussion

In this section, we broadly discuss our approach and the implications and opportunities that it hints towards. We examine the effect of promotion on employee behavior and physiology using mobile sensing consisting of phones, wearables and Bluetooth beacons. We focus on short term effects that are brought about in physical

Figure 3.7: The figures above show the distribution of coefficients of the Logistic Regression model when trained on features obtained from ROCKET. (a) is the distribution of the coefficients when SFS is not performed. We can see that there are quite a number of values that are close to zero. (b) is the distribution of coefficients after performing SFS. The learned coefficients are non-zero, as is expected for important features in the case of logistic regression model.
3.6 Discussion

Detecting Job Promotion Using Mobile Sensing

activity, workplace behavior, wellbeing (stress, sleep), physiology (heart rate) and miscellaneous other features as a result of promotion. Subgroup analysis allowed us to consider gender differences and job performances (viz. high and low performers). Results show that at least within our cohort, promotion leads to reduced stress, more engagement at work and decrease in physical activity. We perform a between-groups analysis to investigate the difference in behavior between promoted participants (after promotion) and non-promoted participants. We also find that some of the effects of promotion are different for males and females. We report the findings with regards to higher and lower job performers and changes in behavior and physiology after promotion.

We believe that our findings present implications and opportunities for future work and applications. First, we show that it is possible to use passive sensing data from mobile devices to dive deeper into workplace promotion and its effects in an objective manner. Although we only concern ourselves with promotion based changes, passive sensing can help us better understand how people react to changes which might not just be limited to promotion (e.g., job changes and role transitions). Understanding the effect of certain job changes or role transitions can help in improving workplace wellbeing. This could be achieved through interventions from management or by simply helping workers keep track of changes in their behavior and life as a result of role change. In addition, learning how a particular person reacts to role changes may help the company in personnel management because then they can take into account the person’s specific situation or behavior in consideration. Management can understand the differences in individual worker’s behavior, physiology and how to help them better deal with things such as changes in communication, getting used to new routines among other things. These are all associated with reacting to change in the job. In contrast, workplace sensing can also offer the ability to study how
workers initiate the change themselves. An interesting avenue that this opens up is with regards to job crafting (345). Job crafting refers to the proactive steps or actions taken by employees utilizing opportunities available to them in order to customize their jobs. For instance, workers can change their tasks (e.g., changing the scope or how they perform it) and interaction with peers (e.g., nature or extent of interactions). Prior research shows that job crafting has positive effects on employee’s psychological wellbeing (37), work engagement and performance (313). With the help of workplace sensing, employers can understand how people with promotion focus craft their job. Equipped with such knowledge, organizations can create appropriate policies to help foster a similar approach for other employees. Prior researchers (183; 312; 50) show that individuals who are focused on being promoted are more likely to craft their job, compared to individuals who are not, as it can lead to status gain, personal growth and success (183; 312; 50). Ultimately, all this information allows management to retain and motivate workers by offering them an environment to grow while at the same time recognizing that there are individual differences in people.

Next, we explore whether we can detect promotion using passive sensing data. We do this via a machine learning based approach. Our pipeline consisting of ROCKET (for feature extraction), Sequential Forward Selection (for feature selection) and Logistic regression (predictive classification model) is able to achieve an AUC of 0.72. However, as a result of performing convolution operation on the time series signals, we lose interpretability of the model. Needless to say, it is complicated trying to interpret how multivariate time signals interplay with each other in order to result in the given performance based on the features obtained after performing the convolution operations. Also, with our classification model we used 100 features as obtained from SFS. But as we can see in figure 3.5, the performance somewhat plateaus for a while after about 20 features until it finally increases again once we hit 80 features.
mark. It might be possible to use some early stoppage criteria during training so that we can stop the training early to reduce the complexity of the model and still end up with similar performance.

Our findings show that passive sensing from mobile devices is capable of detecting promotion based changes in the workplace. In terms of our predictive classification model, it shows that a model trained on passively sensed data of a group of workers can, with some acceptable level of error, detect promotion events for other workers as well. From a practical implementation perspective, we could have, for example, a model pre-trained on a group of worker’s non-promoted and promoted period’s data and use it to detect promotion events for a new worker. Once we have enough data available from a new worker (in our study we use 7 days long time series), we could run it through the model, and it would let us know whether the employee was promoted during that period or not. This is just one of the use-cases of the predictive classification model. We could also envision an online machine learning model that is learning and updating itself in real time settings based on the passive sensing data.

Section 3.7

Privacy for Future of Work

Privacy and security issues need to be considered with the utmost importance as more and more sensing technology is developed for the workplace. New agency programs on future of work (22) are funding groups to study privacy in the context of new technology in the workplace. Information workers use enterprise applications (e.g., a firm’s social media, texting, communications and email systems), video conferences and swipe cards to enter and leave buildings – these are all examples of enterprise technologies that are widely deployed in industries today. As a result, there are a large number of “information breadcrumbs” captured as workers go about their day
to day tasks in the enterprise.

The introduction of new technology in the workplace is typically under the remit of private companies that do not disclose how that data is collected, stored and analysed; or, specifically, how data is use in the case of promotion. An important question is how advanced human sensing technology that observes human behavior (and in our case physiology signals) is going to be safeguarded against abuse and bias of employees. This is a critical question that moves beyond privacy in to workers’ rights. How are workers’ rights to be protected? As ubiquitous computing researchers, we need to consider the implications of advanced sensing technology we help to develop. It is our opinion that workers will only adopt new and potentially invasive sensing technology if they feel they have complete control over their data (e.g., controlling who has access to it) and get something out of the new technology (e.g., health and wellbeing data, performance data to help improve their productivity). These types of sensor driven interventions might be valued by workers as long as they have control of what data is shared and what is private.

How these emerging future of work technologies impact workers’ rights is an open and important area of research with little guidance right now. Many times only employees of the company are aware of what technology is used in the firm. For example, many companies offer health apps with insurance discounts to workers as an option. We understand that our study of using mobile sensing for information workers adds to the burden and pressure on privacy of workers. On the more narrower issue of conducting an ethical study, the participant’s data in our study is kept at a secure, central server of the research institution. The study has an IRB and protocols. There is no personally identifiable information available to the researchers. In addition, workers in our study are notified that their mobile sensing data and surveys responses will not be shared with their employers. While these protections are important for
3.8 Limitations

Detecting Job Promotion Using Mobile Sensing

a study like ours, they are small in consideration to the broader issue of protecting workers data in the era of future of work. In this section, we do not propose specific solutions. Rather, we raise the issue as being critically important to advances in sensing technology in the workplace that must start from the viewpoint of protecting workers’ rights first.

Section 3.8

Limitations

While our research findings are interesting, there are a number of limitations associated with our work. The analysis is based on a small sample of information workers spread throughout different companies in the USA. Therefore, it is difficult to determine the degree to which our results generalize to other contexts, occupations and industries. As such, application of specific findings should be made with caution. In addition, it is possible that the participants in our study had the physiological and behavioral changes because of situations unrelated to promotion. Next, in order to maximize the number of N, we take into account upto 60 days before and after promotion, which is a short duration to examine the effect of life events such as promotion. The results might be entirely different when consideration is put on longer term effects. However, prior studies based on self-reports show that the effect can be long lasting (46; 147; 11). Future studies will consider using objective measures such as passive sensing to investigate if that is still the case when analysis is done on objective data. The proxy for interaction as well as workplace activity that we use in our study is based on the proximity of Bluetooth beacons, which is prone to error. If person A’s phone is detecting person B’s Bluetooth with strong signal strength (and vice versa) then we infer that as an interaction. But, we cannot know for certain whether it was a face to face interaction or simply collocation. Future researchers
on this topic could make use of RFID tags which are more reliable in inferring face to face interactions. In addition, the job performance metrics as well as the promotion ground truths are collected via self-reports which are known to be subject to bias such as recall and social desirability. Furthermore, we do not know if workers report a major or minor promotion. Its entirely up to the workers to decide what they consider to be a promotion and we do not know the specifics of what the promotion entails. Different individuals may interpret promotion differently, although given that we have workers from different companies, a unified way to measure promotion objectively may not be possible. We achieved a good AUC for our predictive classification model utilizing random convolutions. Logistic regression trained on ROCKET based features appears to outperform all the other models. Although using ROCKET based features improve the performance of the model significantly, we end up losing interpretability.

The impact of promotion on an individual has been discussed in prior work. However, so far, self-reported responses have been used to draw insights related to promotion. In this paper, we presented a passive sensing based approach to collect objective data from phones, wearables and Bluetooth beacons in order to explore behavioral and physiological changes of employees after promotion. We utilized up to 60 days of data prior to promotion and 60 days of data after promotion and reported a number of insightful findings. We trained a machine learning model that can identify whether the sensing data belongs to a promoted period or a non-promoted period with an AUC of 0.72. Our work represents the first time mobile sensing has been used to understand the behavioral impact of job promotion on information workers. We believe our
findings pave the way for further research in the future to understand the effects of promotion and other job changes on individuals in the workplace. Understanding employees and how they react to job changes might be useful in order to improve the general wellbeing of the workforce as well as to further future-of-work based research and applications. Finally, we also raised the important issue of securing workers’ rights as new technologies for future of work accelerate.
Chapter 4

MoodCapture: Depression Detection
Using In-the-Wild Smartphone Images

Section 4.1

Introduction

Today, most people automatically unlock their phones using camera biometrics and face recognition. The front-facing camera quietly captures glimpses of users’ faces tens to hundreds of times daily, week in and week out. Unlike selfies, these in-the-moment images capture authentic, unguarded facial expressions, free from biases such as social desirability and self-presentation. We envision a future where AI processes these unguarded facial images on the phone in real-time using deep learning, assessing the user’s mood without needing the images to leave the device, thus safeguarding privacy. This low-burden, continuous approach to depression assessment and detection will significantly alter how mental health is passively assessed, enabling early detection of depression, timely intervention, and constant evaluation of individuals at risk. This
chapter discusses the first steps toward realizing this vision.

Depression is a complex and pervasive mental health issue affecting millions of people worldwide. According to the World Health Organization (WHO), over 264 million people suffer from depression (249), making it a leading cause of disability and a major contributor to the overall global burden of disease. The consequences of depression extend beyond emotional distress (266), significantly impacting physical health (108; 245), social relationships (281), and occupational functioning (85). In severe cases, depression can lead to suicide, accounting for nearly 800,000 deaths each year (9; 39). The need for early detection and intervention in depression is critical, as timely identification of the condition allows individuals to access appropriate treatment and support, thereby improving clinical outcomes and reducing the risk of long-term complications (271; 100). Smartphones offer an opportunity to explore alternative approaches for depression detection that are more objective, unobtrusive, and continuous. The vast amounts of data generated through daily smartphone usage, including images, text messages, and social media interactions, provide a rich and ecologically valid source of information that can be utilized to gain insights into an individuals’ mental state. Consequently, several studies have made use of smartphone sensing data to assess depression (331; 65).

Most of the prior research utilizing facial images to detect depression focus on capturing these images in controlled settings, where individuals may be instructed to perform specific actions (164; 179; 186; 107). These face features are not authentic as they are performative and are influenced by biases such as social desirability and self-presentation. Furthermore, traditional methods such as clinical assessments and subjective self-reports are time-consuming and affected by recall bias. Advances in smartphone cameras offer a solution to address these disadvantages. To this end, we present MoodCapture, a novel approach to collect in-the-wild face images and self-
4.1 Introduction Depression Detection Using Smartphone Images

Figure 4.1: MoodCapture Framework: Users answer the PHQ-8 depression survey questions using the MoodCapture Android App while the app takes bursts of photos using the front-facing camera on the smartphone (top-left). Image characteristics are analysed using factors, such as, illumination, indoor vs. outdoors, phone angle, dominant image color, and background objects (top-right). Given that raw images compromise privacy, these characteristics provide insights into the types of features our machine learning and deep learning model infer. Finally, OpenFace features are extracted to train machine learning models, while raw images are used to train deep learning models (bottom). Depression classification is a binary predictor that classifies an image as depressed or not depressed, whereas PHQ-8 score prediction is a regression model that predicts raw PHQ-8 scores.

reported depression symptoms in natural, everyday environments using smartphones. The resulting face images capture authentic and unguarded facial expressions. Thus, minimizing the influence of self-awareness on emotions and enhancing the credibility of our data. By using such naturalistic images for analysis and training machine and deep learning models, we can better understand intricate patterns associated with depression. Ultimately, insights from our work can be used to create accurate,
efficient, and personalized tools for depression detection.

Our chapter contributes to the growing intersection of Human-Computer Interaction (HCI) research and mental health assessment by investigating the potential of machine learning and deep learning models trained using in-the-wild smartphone images for identifying depressive symptoms. We collected over 125,000 images from N=177 participants diagnosed with major depressive disorder over three months, utilizing 87 distinct types of Android devices owned by users in the study. On average, each participant provided six photos per day, creating a varied and extensive dataset. We comprehensively analyze various image characteristics obtained from these images captured in-the-wild. We evaluate the performance of machine learning and deep learning models trained to predict depression based on these images, as shown in Figure 4.1. At the end of the study period, we assess user acceptance by inquiring about participants’ comfort levels and privacy concerns in sharing their photos for mental health assessment purposes. Therefore, our research aims to foster the development of more ethically sound mental health assessment and intervention tools.

The contributions of our work are as follows:

- We develop a passive-sensing image-based mobile app called MoodCapture that automatically collects in-the-wild smartphone images from participants’ front-facing cameras, ensuring an unobtrusive data collection process and maintaining user privacy. Compared to prior studies, our application captures front-facing photos in-the-moment, resulting in naturalistic images with authentic emotions. Our app provides valuable insights for future in-the-wild studies.

- We analyze different image characteristics such as illumination, location, phone angle, background color, and objects, providing insights into the visual properties of smartphone images. For example, majority of the images were taken indoors in well lit environments. These properties are crucial for model train-
4.1 Introduction Depression Detection Using Smartphone Images

ing and informs HCI practitioners about the environmental conditions in user interactions.

- We evaluate the performance of several machine learning and deep learning models for depression detection and PHQ-8 score prediction. A random forest trained with 3D face landmarks demonstrates the feasibility of analyzing depression from in-the-wild smartphone images, resulting in a balanced accuracy of 0.60, Matthew’s Correlation Coefficient (MCC) of 0.14 and Mean Absolute Error (MAE) of 130.31 (a 6% improvement over baseline on a 0-800 scale). Furthermore, we identify important features providing useful insights for HCI design.

- We report on user acceptance with respect to the comfort levels of the participants in sharing their photos for mental health assessment, providing valuable insights into privacy concerns that inform the future design of in-the-wild image-based mental health assessment tools.

In addition to its relevance to the HCI community, our MoodCapture study contributes to affective computing, which deals with recognizing, interpreting, and simulating human emotions. By leveraging computational methods and machine learning models to interpret emotional cues from images, our research contributes to the understanding and development of affective computing within the HCI field. Furthermore, our study has tangible, real-world implications, such as the potential benefits of early depression detection, timely interventions, improved clinical outcomes, and overall well-being for individuals.

This chapter is structured as follows: Section 4.2, presents related works in depression detection and work that uses smartphone images. Section 4.3 details the MoodCapture study, participant demographics, and the analysis we perform to iden-
4.2 Related Work

In this section, we delve into the pertinent literature, examining the key studies and developments in the field that inform the foundation of our MoodCapture research.

4.2.1. Smartphones and Mental Health

Depression has been traditionally diagnosed through clinical interviews or self-reporting questionnaires such as the Beck Depression Inventory (BDI) (29) and the Hamilton Depression Rating Scale (HDRS) (162). However, these tools are affected by the individuals’ subjective recollections, social desirability bias, mental health stigmas, or the person’s diminished self-awareness (140; 285; 82). Therefore, the pervasive, objective, and continuous nature of multifaceted smartphone data makes it an ideal candidate for unobtrusive depression detection. Many studies evaluate patterns in call logs, text messages, GPS coordinates, and overall smartphone activity, to gain insights into behavioral shifts, social engagement frequencies, and alterations in daily routines, all of which can serve as indicators of deteriorating mental health (331; 349; 65; 234). Other modalities such as speech have also gained traction in evaluating mental health symptoms such as suicidal ideation (33; 255). The growth of social media platforms provides ways to harness user-generated content for depression detection. In particular, analytical approaches using text and images have been applied to content
from platforms like Facebook and Instagram. For instance, the linguistic attributes of posts can shed light on a user’s emotional state, sentiment, and overall mental well-being (81; 61). Moreover, machine learning algorithms have been employed to decipher patterns and indicators of depression from visual content shared on these platforms. Such analyses often encompass aspects like colors, objects, scenes, and overall aesthetics (128; 113; 263).

4.2.2. Contextual Image Factors in Human Computer Interaction

Understanding the content and intrinsic characteristics of spontaneous images could be essential from a HCI standpoint. Contextual elements like environment, angle, color, and lighting play a significant role in how users interact with their smartphones. For example, research by Ikematsu et al. (141) indicates that people often prefer positions that require minimal movement when using their devices. This makes it valuable to examine factors such as the smartphone’s angle and the background objects present during use. In addition, the ambient light during device interaction can act as a situational impairment, as noted by Tigwell et al. (311) and Sarsenbayeva et al. (282). For instance, the facial expressions and illumination on a user’s face can vary greatly between bright outdoor sunlight and controlled indoor lighting conditions. The environment, whether indoor or outdoor, also affects color, which in turn can influence user psychology. Valdez and Mehrabian (324) conducted studies assessing the impact of color on emotions like pleasure, dominance, and arousal. Their findings suggest that colors like blue and purple are typically perceived as pleasant, while greenish hues tend to be more arousing. This raises the possibility that the dominant color in a user’s surroundings might have a correlation with their facial features during smartphone interaction.
4.2 RELATED WORK  DEPRESSION DETECTION USING SMARTPHONE IMAGES

4.2.3. Smartphone Images in Controlled Settings for Mental Health

Extracting facial features to assess mental health and emotions has received significant attention in computer vision, with applications spanning from education to healthcare (215). Here, many studies have explored facial expressions, gaze patterns, and the overall composition of images to extract visual markers symptomatic of depression (164; 179; 186). However, most of these studies are conducted in controlled environments or rely on participants deliberately capturing their images, which could inadvertently influence their emotional portrayal. For instance, Kong et al. (164) captured photographs using a tablet in a standardized clinical setting. Participants were asked to sit before a white background, remove hats or glasses, and tie up long hair to expose their ears; the users looked straight ahead with relaxed expressions as instructed. Similarly, Liu et al. (186) employed a multi-modal deep Convolutional Neural Network (CNN), considering both facial expressions and body movements. During psychotherapy sessions, they captured video using a 4K high-resolution camera in a controlled laboratory setting. Consequently, the participants’ expressions and body movements were analyzed in a highly regulated context. Numerous other studies have similarly relied on advanced devices for image capture, used video recordings, or incorporated additional signals (such as movement, audio) within controlled environments (359; 129; 261; 251; 148; 107).

Our work aims to address these limitations by examining the feasibility of using spontaneously captured images from participants’ smartphones, which offers a more natural and less intrusive method for predicting depression. As smartphones have become an integral part of modern life, they are an ideal tool for unobtrusive and widespread data collection. By utilizing smartphone cameras to capture participants’ images, our approach eliminates the need for controlled environments or deliberate image-taking, thereby reducing the potential for biased emotional portrayals. Fur-
thermore, the widespread availability of smartphones enables our method to reach a larger and more diverse population, ultimately promoting greater accessibility and inclusivity in mental health assessments.

### 4.2.4. “In-the-wild" Smartphone Images for Mental Health

Our study emphasizes the analysis of "in-the-wild" smartphone images, particularly those captured via front-facing cameras of smartphones. These images offer a direct window into an individual's emotions, expressions, and environment, thus enhancing the accuracy of mental health assessments. In contrast to social media content, these images remain relatively free from biases like social desirability and self-presentation, which often affect traditional tools. A limited number of past research have used "in-the-wild" smartphone images for mental health evaluation. For instance, Wang et al. (328) collected 5811 opportunistic photos in-the-wild from 37 students over ten weeks using their phone’s front-facing camera. The study reported that depression scores significantly correlate with the students’ facial expressions and activity. While Wang et al. (328) were the first to use in-the-wild images from front-facing phone cameras to study mental health on a non-clinical population of college students, the authors state that there was insignificant signal in the images to predict self-reported depression. MoodCapture is inspired by this original work, which was part of the StudentLife study (329) in 2013. Our progress is that a decade on from the StudentLife study, phone cameras have seen significant advancements, leading to substantial differences in their capabilities compared to those from ten years ago. For example, new phone cameras typically offer much higher resolution and more megapixels than those from a decade ago, resulting in sharper and more detailed face photos; advances in sensor technology and image processing have greatly improved low-light performance, resulting in today’s phone cameras capturing better quality face photos in low-light conditions; optical image stabilization has become more common in smartphone cam-
eras today, reducing the impact of shaky hands and resulting in smoother sharper photos, especially in low light; and finally front-facing cameras primarily designed for selfie shots have improved significantly in terms of resolution, image quality, autofocus on the face. Other differences between Wang et al. (328) and our work are that we take advantage of massive advances presented by deep learning models and focus not on a non-clinical group but a clinical population.

Other studies have also leveraged front facing cameras in one way or another. Khamis et al. (152) studied the visibility of the face and eye in 25,726 in-the-wild images of smartphone users and found that the full face is visible about 29% of the time. The authors stated that their state-of-the-art face detection algorithm performed poorly against photos taken from front-facing cameras. Similarly, Bâce et al. (19) used in-the-wild images to study the visual attention and gaze of users. Darvari et al. (78), on the other hand, used in-the-wild images from rear-facing cameras. The authors developed a smartphone application that allows users to periodically log their emotional state together with pictures from their everyday lives. They collected 3,305 mood reports with photos from 22 participants. Authors report finding context-dependent associations between objects surrounding individuals and their self-reported emotional state. However, the genuine spontaneity of these captures and their potential for unbiased mental health evaluation remain relatively unexplored. Our contribution to this growing field pivots on the innovative use of genuinely spontaneous, in-the-wild facial images for depression detection. By employing a passive-sensing mobile application that seamlessly captures images without the subject’s acute awareness, we negate the potential influence of self-awareness on emotional representation. This strategy bolsters the ecological validity of our data source, making it a robust tool for depression detection.
4.3 Methodology

In what follows, we discuss the design of our MoodCapture study, demographic information of the individuals that participated in the study and the ground-truth used for analysis.

4.3.1. Study Design

We recruited 181 participants from across the United States using targeted online advertisements on Google and Facebook. Each participant underwent a clinician-administered Structured Clinical Interview for DSM-5 (SCID), and only those diagnosed with Major Depressive Disorder (MDD), without bipolar disorder, active suicidality, or psychosis, were eligible for the study. Upon qualification, participants installed our Android-based mobile sensing app on their devices, which gathered Ecological Momentary Assessments (EMA) during the 90-day study period. The app prompted participants to complete a brief Patient Health Questionnaire-8 (PHQ-8) (171) (see Table A.1) survey about their depressive symptoms three times daily (morning, afternoon, and evening). As participants answered their daily surveys, the app was designed to discreetly capture a burst of up to 5 images using the front-facing camera. Specifically, images were taken when participants responded to the PHQ-8 item: "I have felt down, depressed, or hopeless." (see Fig. 4.2). We chose this question as we believed it would best capture participants’ genuine emotions related to depression. The PHQ-8 is a validated inventory for measuring depression. For further information about the survey, please refer to the Ground Truth section.

During the onboarding process, we informed participants about the image capture procedure and emphasized that sharing their photos was optional. Upon launching the mobile app for the first time, participants were asked, "To help us better understand
your depressive symptoms, we would like to take a few photos in the background that capture your facial expressions while you fill out questionnaires. Do you give us permission to do this?” Participants could respond with either "Yes" or "No." If they agreed to share their photos, the app captured images as they answered the EMA. If they opted not to share their photos, no images were captured. The image capture process was designed to be unobtrusive, with only a green dot at the top of the Android status bar/screen indicating camera usage – which users’ may or may not have observed. Participants did not see their face or receive any other indication that photos were being taken. This discreet image capture process ensured a seamless user experience without interrupting or obstructing the EMA flow. As stated earlier; while participants consented to have photos taken using the front-facing camera during the operation of the MoodCapture app in the study they were not informed exactly when these photos were captured, thus promoting in-the-moment naturalistic and authentic capture of users’ faces and surroundings.

Participants were compensated $1 for each completed EMA, with an additional $50 bonus for achieving a completion rate of 90% or higher during the study period. Compensation was not dependent on sharing photos; participants were compensated regardless of their photo consent. The study was approved by Dartmouth College’s Internal Review Board (IRB). Our analysis and predictive modeling focuses on 177 out of the 181 participants who provided consent for their photos to be captured. We collected 125,335 images from these participants, excluding 15,063 photos that were either too blurry, contained no faces, featured children, or contained nudity.

4.3.2. Demographics

The majority of participants in our study identified as female (86.4%, N=153) followed by male (9.6%, N=17) and non-binary (2.8%, N=5). In terms of race, 83.6% (N=148) are White, 2.8% (N=5) are Asians, 4.5% (N=8) are Black or African American, 0.5%
4.3 Methodology: Depression Detection Using Smartphone Images

Figure 4.2: PHQ-8 application screens for each item: Images are always captured while users respond to the PHQ-8 depression survey question (highlighted in cyan): “I have felt down, depressed, or hopeless”. While users consent to have photos taken using the front-facing camera during the operation of the MoodCapture app they are not informed exactly when these photos are captured to promote in the moment naturalistic and authentic images.
(N=1) are American Indian/Alaska Native and 6.7% (N=12) belong to more than one race. See Table 4.1 for the detailed breakdown.

Table 4.1: Demographics, smartphones, and image composition in our study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>153</td>
<td>86.4%</td>
</tr>
<tr>
<td>Male</td>
<td>17</td>
<td>9.6%</td>
</tr>
<tr>
<td>Non-binary</td>
<td>5</td>
<td>2.8%</td>
</tr>
<tr>
<td>Other (prefer to self-describe)</td>
<td>2</td>
<td>1.1%</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>148</td>
<td>83.6%</td>
</tr>
<tr>
<td>Asian</td>
<td>5</td>
<td>2.8%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>8</td>
<td>4.5%</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>More than one race</td>
<td>12</td>
<td>6.7%</td>
</tr>
<tr>
<td>Other (prefer to self-describe)</td>
<td>3</td>
<td>1.6%</td>
</tr>
<tr>
<td><strong>Smartphones</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samsung</td>
<td>107</td>
<td>60.4%</td>
</tr>
<tr>
<td>Google</td>
<td>36</td>
<td>20.3%</td>
</tr>
<tr>
<td>Motorola</td>
<td>19</td>
<td>10.7%</td>
</tr>
<tr>
<td>Other</td>
<td>15</td>
<td>8.4%</td>
</tr>
<tr>
<td><strong>Image Resolution</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3648x2736</td>
<td>57</td>
<td>32.2%</td>
</tr>
<tr>
<td>3264x2448</td>
<td>52</td>
<td>29.3%</td>
</tr>
<tr>
<td>2640x1980</td>
<td>16</td>
<td>9.0%</td>
</tr>
<tr>
<td>Other</td>
<td>52</td>
<td>29.3%</td>
</tr>
</tbody>
</table>

4.3.3. Ground Truth

Our study is designed to account for the wide variability in MDD symptoms. In particular, MDD can manifest in over 1000 distinct symptom combinations across individuals, with significant within-day variations (73; 111; 94; 110). However, existing diagnostic methods face several limitations. Firstly, SCIDs are not effective in capturing moment-to-moment fluctuations in depression symptoms. Secondly, the Likert scale used in depression screening tools like the PHQ-8, which typically offers a limited response range from 0-3, forces respondents to fit their experiences into
4.3 Methodology Depression Detection Using Smartphone Images

pre-set categories. This can lead to central tendency bias and a lack of detailed responses for complex mental states. To overcome these challenges and better capture intra-individual variation, our clinical team modified the PHQ-8 scale to a more nuanced continuous scale ranging from 0-100 (see Figure 4.2). The practice of re-scaling psychometric scales is not uncommon and has been applied to the PHQ in various past studies (196; 236; 127; 191). A standard PHQ-8 score of 10 or higher (out of 24) signifies major depression (180). In our continuous scale, a score exceeding 334 indicates depression (i.e., 10/24 times 800). To provide holistic analysis, we complement our binary classification models with regression models that predict raw PHQ-8 scores. Note that the PHQ is versatile, serving both as a screening tool for depression and as a means to monitor clinical symptom changes (169).

To enhance the reliability and accuracy of the EMA responses, we employed a validation technique wherein the app randomly reversed one question in each PHQ-8 survey (thus adding an additional item), ensuring that participants are attentive. We then compared the responses to the original and reversed questions; if there is a significant discrepancy, the response is excluded from our analysis. After applying this filtering process, we obtain a refined dataset comprising 31,215 EMAs. Since we captured a burst of images with each EMA response, we amassed 110,272 images in total. As depicted in Figure 4.3a, we divided our dataset into two groups: depressed (74,347 images, N=175) and non-depressed (35,925 images, N=156). On average, participants submitted 176 EMAs (stdev = 78) and 623 images (stddev = 278) per participant during the study period. It is crucial to note that all participants recruited for this study had major depressive disorder. Consequently, they reported being below the cut-off threshold on some days and above it on others. However, 19 participants consistently reported depression throughout the study.

Figure 4.3b shows the variability of PHQ-8 scores among participants i.e., intra-
4.3 Methodology Depression Detection Using Smartphone Images

(a) Distribution of PHQ-8 Scores

(b) Intra-individual variability

Figure 4.3: PHQ-8 score statistics: Figure (a) depicts the distribution of the PHQ-8 score reported by the participant and the corresponding label (i.e., Depression or No Depression). Figure (b) shows the variability of PHQ-8 scores among participants over the duration of the study (Cronbach’s $\alpha = 0.85$).

individual variability. It provides insight into the fluctuations in a participant’s scores over time. On average, participants’ scores varied around their own mean by approximately 101.92 points, with the variability ranging widely from a standard deviation of 27.56 points to as high as 262.24 points. This suggests that some participants had relatively stable scores over time, while others exhibited more pronounced fluctuations. Moreover, we measured the internal consistency of the PHQ-8 items, obtaining a Cronbach’s $\alpha = 0.85$. This demonstrates good reliability and validity of our measures.

4.3.4. Image Characteristics

We gather in-the-wild images captured by participants using a diverse range of smartphones with varied configurations and camera placements. Predominantly, participants use Samsung, Google, and Motorola devices, and the images captured from these devices had resolutions ranging from 1920x1080 to 4656x3488 (see Table 4.1). Our naturalistic approach at capturing image ensures ecological validity and represents users’ natural behavior while engaging with their devices in different environments. To examine the characteristics of these images, we analyze factors such as phone angle, dominant color, lighting condition, photo location, and background.
elements present in the photos. The in-the-wild smartphone images offer a unique
glimpse into the multitude of ways users interact with their devices and surround-
ings. However, extracting meaningful insights from these images demands a refined
approach that acknowledges the diverse contexts in which they are captured. To
achieve this, we utilize the BLIP (181) visual question answering (VQA) model, an
advanced AI tool specifically designed for image analysis and answering questions
about image content and context. BLIP is recognized as a state-of-the-art method
for visual question answering tasks. Furthermore, the VQA analysis contextualizes
our predictive modeling in the following ways. First, it can elucidate the raw image
content, which is the input for our deep learning models. Second, as our ML mod-
els use handcrafted features from the face, it differentiates the performance obtained
by considering background in addition to face versus only face. In summary, our
motivation is to harness VQA to interpret both explicit and implicit image content.
Consequently, enabling a more holistic approach to image analysis, where both the
central subjects and their surrounding context contribute to the predictive insights.
Importantly, as we cannot display images to protect participant privacy, the VQA
provides some level of interpretation. With the help of the VQA model, we explore
the following characteristics:

**Image Angle:** By inquiring about the image angle, we gain an understanding of user
interaction dynamics with their devices. Varying angles, such as high or low, offer
insights into users’ physical engagement with their smartphones. High, low, or level
angle refers to the perspective from which an image is captured or taken with respect
to the subject in the frame. A low angle shot refers to the subject looking down at
their phone, whereas a high angle shot refers to the user looking up at their phone.
A level angle shot is taken from the same height as the subject, capturing it at eye
level. We asked the VQA: “Is the image taken from a high, low, or level angle?”.
4.3 Methodology

**Dominant Colors:** Colors are crucial for establishing the context of an image. To identify dominant colors in the images and understand the users’ environments, we asked the VQA: “What is the dominant color of the image?”

**Lighting Condition:** Lighting conditions in an image reveal important information about the user’s ambient environment. Using the VQA model, we classified images based on their lighting as well-lit, dimly lit, or poorly lit. We asked the VQA: “Is the image well-lit, dimly lit, or poorly lit?”

**Photo Location:** The location context (indoors or outdoors) can significantly influence user-device interactions. We determined the location context of images with the help of the VQA model by asking: “Is the photo taken indoors or outdoors?”

**Background Objects:** Identifying specific objects in the background can provide valuable information about the user’s context and activities. We queried the VQA model about the background objects to recognize and categorize various elements within the images. We asked the VQA: “What are the background objects in the photo?”

**Number of People in the Image:** In order to evaluate the social context of the images, we employed the VQA model to determine the number of people present in each image. This information provides insight into users’ social interactions and their surroundings during device usage. We asked: “How many people are in the image?”

By leveraging the BLIP VQA model, we are able to extract structured insights about the content and context of in-the-wild images, enhancing our understanding of user behavior and interaction with their devices in diverse settings. Importantly, two expert annotators manually annotated 1500 unique images corresponding to individual EMAs. To clarify any ambiguities, we provided them with specific instructions. They determined the image angle based on eye level with the phone. ‘Dominant color’ refers
4.3 Methodology  Depression Detection Using Smartphone Images

to the most prominent color in the overall image. For lighting conditions, ‘well-lit’ represents the best lighting condition, while ‘poorly lit’ indicates the worst. After completing the manual annotation, we calculated the average accuracy between the two annotators and the inter-rater agreement using Cohen’s kappa ($\kappa$). These results (see Table 4.3) indicate substantial agreement between the annotators and alignment with VQA responses, indicating high reliability, consistency and accuracy.

4.3.5. Depression Classification and Regression

In this study, we aim to accurately identify depression from facial images by utilizing both machine learning and deep learning techniques. In particular, we build binary classification models to classify a face image as depressed or not depressed, and a regression model to predict the raw PHQ-8 score (see Section 4.3.3).

**Machine Learning.** To facilitate machine learning approaches, we extract 711 (709 trainable) facial features using OpenFace (25), a well-validated feature set for depression detection that has been employed in a variety of studies (120; 274; 252). The extracted features consist of 2D and 3D facial landmarks, head pose, eye gaze, facial expressions represented by facial action units (FAU), and rigid and non-rigid shape parameters (see Table 4.2). Before training, we apply feature selection using only the training set in two distinct ways. First, we compute the mutual information (MI) metric, selecting the most independent features indicated by smaller MI values. In our analysis, we choose the top 25%, 50%, or 100% of the features. Second, we conduct an ablation study to gain valuable insights into the effectiveness of different handcrafted features, thus inferring the best performing feature set. An ablation study is a systematic experimental procedure in which certain features are systematically removed or “ablated” to analyze their individual impact on the overall model performance. We use a Logistic Regression (135) model for classification and an ElasticNet
4.3 Methodology  Depression Detection Using Smartphone Images

Table 4.2: Description of OpenFace features (25).

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAU (23)</td>
<td>35</td>
<td>FAU uses the facial action coding system to describe anatomically possible expressions resulting from muscle activations. It indicates the presence and intensity of an expression(^1).</td>
</tr>
<tr>
<td>Gaze (344)</td>
<td>8</td>
<td>Direction vector in eye gaze direction as measured from the eye location, consisting of location and angle. This feature is computed after landmark detection.</td>
</tr>
<tr>
<td>Eye Landmarks</td>
<td>280</td>
<td>Contains 2D and 3D landmarks describing various positions of the eye.</td>
</tr>
<tr>
<td>Head Pose</td>
<td>6</td>
<td>Describes translation in millimeters with respect to camera centre and rotation in radians around x, y, and z axes.</td>
</tr>
<tr>
<td>Rigidity Parameters</td>
<td>40</td>
<td>These parameters are divided into rigid and non-rigid shape parameters. Rigid shape parameters describe the face’s positioning within an image that includes aspects like scale, rotation, and movement. Conversely, the non-rigid shape parameters focus on the variations in facial appearance caused by individual characteristics or expressions, such as variations in facial width or height, smiles, blinking, and other facial expressions.</td>
</tr>
<tr>
<td>2D Landmarks (24)</td>
<td>136</td>
<td>These are x and y axes locations of different face landmarks in the image. These landmarks refer to specific locations in the face. For example, a point in the right eye is represented as landmark number 38, while points in the lips are represented using numbers 49-68. All landmark numbers are described in (276; 275).</td>
</tr>
<tr>
<td>3D Landmarks (24)</td>
<td>204</td>
<td>These are x, y, and z axes locations of different face landmarks in the image. The landmark numbers are identical to 2D landmarks, however, they are represented using three coordinates.</td>
</tr>
</tbody>
</table>

(360) for our regression task, whereas a Random Forest (RF) (49) is used for both tasks. Statistical approaches such as regression and a bagging-based decision tree can
provide different modeling insights. The baseline model is a RF trained using the participant’s gender, age, and time spent on EMA.

**Deep Learning.** Deep learning models are capable of learning useful features directly from raw images. Pre-trained computer vision models trained on large-scale datasets can capture image features that are transferable to other domains. As a result, we examine the performance of various EfficientNet (308) and InceptionResNetv3 variants, which were previously trained on the ImageNet and VGGFace2 datasets, respectively. Upon observing that the EfficientNet B0 (EffNet) model provided the best performance while other models were underfitting our dataset, we decided to further fine-tune EffNet for depression prediction. We implement EffNet using the PyTorch framework, freezing all layers during the training process except for blocks 6 and 7. The classification and regression models are fine-tuned using binary cross-entropy and mean absolute error loss functions, respectively. For optimization, we use the Adam optimizer (with a learning rate of 0.0001) with a batch size of 256 trained for 50 epochs. This fine-tuning process allows the model to learn and adapt to the specific characteristics of our depression detection dataset, potentially improving its performance and generalizability.

**Evaluation.** To effectively evaluate our models, we adopt a 5-fold leave-subject-out cross-validation approach. This method ensures that all images associated with a single participant are exclusively used for training, validation, or testing the model but not mixed among the subsets. Furthermore, we use nested cross-validation on our training data for hyper-parameter tuning. The subject-independent splits and cross-validation ensure our results are more robust than those of a single train-test strategy. We evaluate classification performance using balanced accuracy (Equation 4.1) and Matthew’s Correlation Coefficient (MCC) (Equation 4.2), whereas regression
performance is evaluated using MAE (Equation 4.3). We chose these metrics as they provide a comprehensive assessment. For example, MCC summarizes all four values in the confusion matrix, whereas balanced accuracy emphasizes both true positive and true negative detection. In fact, MCC is preferred over F1 score in many binary classification problems (64).

\[
\text{Balanced Accuracy} = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \tag{4.1}
\]

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.2}
\]

where \(TP, TN, FP,\) and \(FN\) are true positives, true negatives, false positives, and false negatives, respectively. Note that higher balanced accuracy and MCC values indicate better performance. MCC ranges from -1 to +1, where +1, 0 and -1 indicate perfect classification, random coin-toss classification, and perfect mis-classification, respectively. The regression models are evaluated using MAE defined as:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{4.3}
\]

where \(N\) is the number of samples. \(y_i\) and \(\hat{y}_i\) are true and predicted PHQ-8 scores, respectively. Note that lower MAE values indicate better performance.

\section*{Results}

In this section, we present the outcomes of our analysis, which includes an examination of image characteristics, an evaluation of the predictive capabilities of our machine learning models, and an ablation study. In addition, we identify crucial features
4.4 Results Depression Detection Using Smartphone Images

integral to our models’ performance and explore potential biases within these models.

4.4.1. Image Characteristics

Our analysis using the VQA model reveal many insights into different features of real-world smartphone images. These images serve as glimpses into user interactions and surroundings. From Table 4.3 and Figure 4.4, we notice that the VQA obtained good accuracy ranging from 89% for lighting conditions to 97% for number of people in the image. Furthermore, $\kappa$ is greater than 0.70 for all questions indicating substantial inter-rater agreement. In terms of capture angle, the images predominantly favored a low angle, with approximately 96.08% falling into this category. Conversely, a mere 3.92% were captured from a high angle, suggesting a specific user posture or device interaction habit in the majority of instances. Dissecting the dominant colors present, we found that ‘white’ emerged as the prevailing color, characterizing roughly 67.51% of the dataset. Other noticeable colors included ‘black’ at 8.70%, while a combined representation of ‘brown’, ‘blue’, ‘gray’, and ‘yellow’ accounted for approximately 18%. A diverse array of other hues constituted the remaining 5.75%, emphasizing the richness of user environments. Closer analysis during the annotation process revealed that the images’ dominant white color mainly reflects environmental elements like white walls and ceilings, not participants’ skin tones. Importantly, we noticed most images consisted of partial face images, an observation commonly found in other similar studies (152). Hence, the dominant color is influenced by background objects. This is evidenced in Figure 4.4, where walls, ceilings, tiles and lights are frequently identified as background objects, ensuring our analysis focuses on environmental, not physiological, aspects. The lighting conditions under which these images were taken were also revealing. A vast majority (80.57%) were captured under well-lit conditions, indicating optimal settings for smartphone interaction. The dimly lit and poorly lit categories followed with 10.35% and 9.08%, respectively, showcasing the
4.4 Results

Depression Detection Using Smartphone Images

Figure 4.4: Background objects: Word Cloud showing the range of objects detected in the background of the images captured. (Acc=91.72; \( \kappa = 0.70 \))

varied ambient conditions in which users interact with their devices. Furthermore, in terms of photo location, an impressive 95.08% of the images were taken indoors, signifying the primary environment for user-device interaction. The outdoor segment, constituting 4.92%, provided insight into the more dynamic and mobile interactions users might experience. Notably, 95.81% of the captured images featured only one person. Regarding background objects, we discovered that walls, lights, pictures, and windows were the most common elements. The presence of terms such as "pillow" could imply individuals reclining, while words like "plant," "moon," "flower," and "cloud" might suggest outdoor settings. Overall, it appears that a significant number of images were captured indoors against plain backdrops, possibly within homes or offices. To visually represent these background objects, we have created a word cloud, which can be seen in Figure 4.4.

4.4.2. Predictive Analysis

In our analysis, we leveraged both machine learning and deep learning to assess MoodCapture’s ability to detect depression in natural settings. As shown in Table
Table 4.3: Image Characteristics: Different characteristics of the image captured, such as image angle, dominant colors, lighting conditions, photo location and number of people present. The accuracy and Cohen’s kappa are presented in braces next to the categories. These results indicate substantial agreement between the annotators and alignment with VQA responses, indicating high reliability, consistency and accuracy.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image Angle</strong> (Acc=96.67; ( \kappa =0.82 ))</td>
<td></td>
</tr>
<tr>
<td>Low angle</td>
<td>105,949 (96.08%)</td>
</tr>
<tr>
<td>High angle</td>
<td>4,323 (3.92%)</td>
</tr>
<tr>
<td><strong>Dominant Colors</strong> (Acc=96.81; ( \kappa =0.75 ))</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>744,14 (67.51%)</td>
</tr>
<tr>
<td>Black</td>
<td>9,586 (8.70%)</td>
</tr>
<tr>
<td>Brown</td>
<td>6,053 (5.49%)</td>
</tr>
<tr>
<td>Blue</td>
<td>5,809 (5.27%)</td>
</tr>
<tr>
<td>Gray</td>
<td>5,197 (4.72%)</td>
</tr>
<tr>
<td>Other</td>
<td>9,213 (8.31%)</td>
</tr>
<tr>
<td><strong>Lighting Conditions</strong> (Acc=89.03; ( \kappa =0.81 ))</td>
<td></td>
</tr>
<tr>
<td>Well lit</td>
<td>88,843 (80.57%)</td>
</tr>
<tr>
<td>Dimly lit</td>
<td>11,418 (10.35%)</td>
</tr>
<tr>
<td>Poorly lit</td>
<td>10,011 (9.08%)</td>
</tr>
<tr>
<td><strong>Photo Location</strong> (Acc=98.27; ( \kappa =0.71 ))</td>
<td></td>
</tr>
<tr>
<td>Indoors</td>
<td>104,800 (95.08%)</td>
</tr>
<tr>
<td>Outdoors</td>
<td>5,472 (4.92%)</td>
</tr>
<tr>
<td><strong>No. of People in the Image</strong> (Acc=97.89; ( \kappa =0.75 ))</td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>105,657 (95.81%)</td>
</tr>
<tr>
<td>Two</td>
<td>523 (0.47%)</td>
</tr>
<tr>
<td>Three +</td>
<td>8 (0.01%)</td>
</tr>
<tr>
<td>None</td>
<td>4084 (3.71%)</td>
</tr>
</tbody>
</table>

4.4, the EffNet model shows better performance in correctly identifying classes, as evidenced by its 0.61 balanced accuracy. However, it is interesting to note that the RF model outperforms the deep learning model in terms of overall classification (MCC of 0.14) and regression task i.e., predicting PHQ-8 scores (with an MAE of 130.31). Notably, RF achieves a lower MAE than the baseline model (130.31 vs. 138.18), indicating an approximately 6% improvement. This makes it more robust for handling measurement errors when setting ground truth thresholds for classification.
Table 4.4: Performance: Depression detection using machine learning and deep learning methods. Standard deviation is given in braces. ‘LR + EN’ refers to logistic regression for depression classification and elastic net for regression i.e., raw PHQ-8 score prediction. $R^2$ values are presented in Appendix A.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Balanced Accuracy</th>
<th>MCC</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.52 (0.05)</td>
<td>0.03 (0.11)</td>
<td>138.18 (3.71)</td>
</tr>
<tr>
<td>0-3 LR + EN (MI)</td>
<td>0.52 (0.02)</td>
<td>0.04 (0.03)</td>
<td>135.40 (2.65)</td>
</tr>
<tr>
<td>Random Forest (MI)</td>
<td>0.54 (0.02)</td>
<td>0.06 (0.04)</td>
<td>134.45 (2.05)</td>
</tr>
<tr>
<td>Random Forest (3D Landmarks)</td>
<td>0.60 (0.04)</td>
<td><strong>0.14 (0.08)</strong></td>
<td><strong>130.31 (3.94)</strong></td>
</tr>
<tr>
<td>EffNet</td>
<td><strong>0.61 (0.02)</strong></td>
<td>0.03 (0.00)</td>
<td>137.19 (3.67)</td>
</tr>
</tbody>
</table>

Furthermore, we gain several modeling insights from Table 4.4. First, we observe that RF outperforms LR across all metrics, suggesting that decision trees with bagging are useful in modelling face features for depression. RF’s ability to model non-linear dependencies and in-built feature selection makes it a good candidate for our problem. Second, we notice that manual feature selection, such as using 3D landmarks offer better performance than using automatic feature selection methods with MI. This finding underscores the importance of conducting an ablation study to determine the most impactful features for our analysis.

In summary, an RF trained with 3D landmarks performs well across both classification and regression tasks indicated by well-balanced scores across balanced accuracy (0.60), MCC (0.14), and MAE (130.31). Moreover, RF offers better explainability compared to deep learning methods, making it an ideal choice for post-hoc analysis (Section 4.4.4). Our investigation into depression detection and PHQ-8 prediction using machine learning and deep learning methods provides important insights into the potential of different techniques when applied to MoodCapture data in naturalistic conditions. The results emphasize the importance of considering a range of methods, from deep learning models capable of learning complex features to traditional machine learning techniques that offer interpretability and simplicity. By carefully selecting and fine-tuning these models, we can improve the overall performance and
applicability of depression detection systems in real-world scenarios.

4.4.3. Ablation Study

In this analysis, we aimed to determine if specific OpenFace feature sets are more useful for depression detection by evaluating the performance across the seven groups (Facial action units, Gaze, Eye landmarks, Pose, Rigidity Parameters, 2D and 3D landmarks). From Table 4.5, we make several interesting observations that provide insights into the utility of individual feature sets.

First, we notice that many feature sets perform better than the automatic feature selection using MI, indicating that only some specific features in the image are useful for depression detection. This finding suggests that a more focused approach to feature extraction and selection may improve overall performance. Second, we observe that facial action units are less discriminative than other features. This result may be attributed to the presence of partial face images, which are common in front-facing cameras, thus hindering the effectiveness of action units in detecting depression. Third, we find that gaze features outperform eye landmarks, suggesting that gaze direction and angle are useful. These observations highlight the importance of capturing subtle facial changes when developing depression detection systems.

Table 4.5: Ablation Study: Investigating depression detection of OpenFace feature sets using a random forest. The standard deviation is presented in braces. $R^2$ values are presented in Appendix A.2.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Balanced Accuracy</th>
<th>MCC</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial Action Units</td>
<td>0.54 (0.02)</td>
<td>0.06 (0.04)</td>
<td>133.34 (3.02)</td>
</tr>
<tr>
<td>Gaze</td>
<td>0.55 (0.02)</td>
<td>0.10 (0.05)</td>
<td>132.57 (3.11)</td>
</tr>
<tr>
<td>Eye Landmarks</td>
<td>0.54 (0.02)</td>
<td>0.08 (0.04)</td>
<td>132.04 (2.90)</td>
</tr>
<tr>
<td>Head Pose</td>
<td>0.55 (0.03)</td>
<td>0.11 (0.05)</td>
<td>131.01 (3.68)</td>
</tr>
<tr>
<td>Rigidity Parameters</td>
<td>0.55 (0.05)</td>
<td>0.06 (0.04)</td>
<td>133.11 (3.08)</td>
</tr>
<tr>
<td>2D Landmarks</td>
<td>0.53 (0.03)</td>
<td>0.05 (0.06)</td>
<td>132.62 (3.41)</td>
</tr>
<tr>
<td>3D Landmarks</td>
<td>0.60 (0.04)</td>
<td>0.14 (0.08)</td>
<td>130.31 (3.94)</td>
</tr>
</tbody>
</table>

Table 4.5 also indicates that 3D Landmarks is the best performing feature set.
4.4 Results

Depression Detection Using Smartphone Images

for classification and PHQ-8 prediction across all metrics (balanced accuracy=0.60; MCC=0.14; MAE=130.31). These results suggest that an RF trained with 3D landmarks is more accurate, correlates better with the ground truth, and has lower PHQ-8 prediction errors than other methods. 3D landmarks (see Table 4.2) are coordinates of specific points on the face. For example, a point in the right eye is represented as landmark number 38. This location is represented using coordinates. All landmark numbers are described in (276; 275). Intuitively, different values of 3D landmarks correspond to changes in facial expressions over time.

In conclusion, the ablation study provides valuable insights into the utility of specific feature sets for depression detection. By understanding the strengths and limitations of individual features, researchers and practitioners can make informed decisions when designing and implementing depression detection systems, ultimately improving overall performance and applicability in real-world scenarios.

4.4.4. Machine Learning Feature Importance

It is crucial to understand important face features that are correlated with depression. Therefore, we employ a post-hoc explainability approach, namely SHapley Additive exPlanations (SHAP) (192), to investigate our best performing (Table 4.4) Random Forest model. SHAP explains the model outputs using notions from game theory. It assigns each feature an importance value for a particular prediction, offering insights into how and why a model makes its decisions.

The top ten important features for depression classification and regression are shown in Figure 4.5. Here, we observe that lips and face contour position are useful for both depression classification and score prediction. For instance, we notice that larger values of face contour near the left cheek (X_14, X_13, X_11) influence the model towards predicting depression and push the raw PHQ-8 score higher. Interestingly, we find that important eye and lip features occur on the right side of the face (Y_48,
4.4 Results Depression Detection Using Smartphone Images

(a) Important features for depression classification.

(b) Important features for predicting raw depression score.

Figure 4.5: SHAP plots describing the top 10 features for the classification and regression tasks. The best performing random forest trained using 3D landmark features is evaluated using SHAP. The features are x and y axis with the numbers (0-indexed) corresponding to facial landmarks (276; 275).

Y_36, Y_17, Y_41); and higher values are associated with higher depression scores. This indicates that the ML model captures asymmetry associated with front-facing camera pictures, i.e., the right side of the face could be more visible. We discuss this further in Section 4.6.

4.4.5. Investigating Bias in Machine Learning

Our dataset predominantly consists of white females, highlighting the need to assess biases in our machine learning models related to gender and race. Thus, we categorize the test dataset into two groups for gender: females, and a combined group of males and non-binary individuals. We made the decision to combine the groups due to the notably smaller representation of non-binary individuals and males in our study. This decision aimed to address the imbalance and ensure a more meaningful analysis, acknowledging the constraints posed by the limited sample sizes of these specific demographic groups. Similarly, we classify the data into white and non-white categories for race. Again, this binary grouping strategy is designed to increase group sizes, thereby improving the statistical power of our analysis. We use our best per-
4.4 Results Depression Detection Using Smartphone Images

(a) Gender-wise performance comparison

(b) Race-wise performance comparison

Figure 4.6: Random forest performance on sub-populations divided by gender and race. Note that balanced accuracy and MCC are multiplied by 100 for better visualization.

The analysis of these biases is intended to enhance transparency in machine learning models, providing insights for future research in this area.
4.5 Ethics

Depression Detection Using Smartphone Images

Section 4.5

Ethical Considerations and User Acceptance

Figure 4.7: Comfort Level: Participant’s comfort with the automated capture of their photos.

In studies involving sensitive mental health data, it is paramount to address the ethical implications to safeguard participants’ privacy, confidentiality, and well-being. Our primary goal was to prioritize the security and confidentiality of the data. We securely stored all collected data and granted access only to specific team members. We took great care in removing all personally identifiable information by implementing a thorough anonymization process. To respect privacy, any image that unintentionally captured subjects or nudity was identified during a review by two team members and subsequently deleted. We understand the sensitive nature of mental health and made sure to maintain transparency with our participants. They were informed about
the study’s purpose, methodology, and expected outcomes. This approach not only sought their permission but also ensured they felt comfortable and safe throughout the process. We further clarified that their compensation was unrelated to their photos.

At the end of the study, we asked participants about their comfort levels with automated front-facing photo capture during surveys. This was optional, so we have responses from only 172 out of the 181 participants that were recruited. Approximately 45% of participants were comfortable, while 38% felt it was intrusive or uneasy, and the remaining 17% were neutral. If participants were uncomfortable, we further ask them about specific reasons for their feelings which can be summarized into a few key themes, as shown below. While we acknowledge these concerns, it is important to note that the study followed strict privacy and data protection guidelines.

(a) **Privacy and Surveillance:** Participants felt uncomfortable with the idea of being watched or monitored, as it evoked a sense of intrusion into their personal space. One participant mentioned, “I don’t like being watched. I’m already paranoid when it comes to cameras.”

(b) **Appearance and Self-Esteem:** Several participants mentioned their discomfort with having their photos taken due to concerns about their appearance. One participant stated, “I don’t want people to see photos of me”, while another said, “I am very uncomfortable with my appearance when I’m depressed.”

(c) **Inappropriate Situations:** Participants worried about the possibility of photo bursts being taken during inconvenient or inappropriate moments. One participant shared, “If I was comfortable and at home, during some of them I may not have been completely covered.”

(d) **Data Security:** Although participants were aware of the study’s data protection measures, some still expressed concerns about the safety and storage
of their images. One participant expressed, “The idea of my picture being out there ...although I know it was to be analyzed with AI.”

(e) **Lack of Control:** Participants felt uneasy about not being able to review, approve, or delete the photos taken during the bursts, as well as not knowing when the camera was active. A participant shared, “Having pictures taken and not knowing what they looked like or if they were embarrassing is an uncomfortable thing to think about.”

In summary, participants’ concerns mainly revolved around privacy, self-esteem, potential inappropriate situations, data security, and control over the images. It is essential to consider these concerns when designing and implementing studies involving photo bursts or similar data collection methods to ensure participants’ comfort and trust in the research process. Acknowledging the sensitive nature of our research, we offered participants the option to delete their photos at the end of the study if they felt uncomfortable. Interestingly, no participants chose this option, highlighting the trust they placed in our research process and commitment to ethical conduct. We remain keenly aware of the potential for technology misuse, especially in unauthorized surveillance or data mining scenarios. We have taken measures to minimize such risks, emphasizing that our technological developments are primarily intended as health aids, not tools for unwarranted monitoring. Further, to address participants’ concerns regarding privacy and data security, one possible solution could be leveraging the capabilities of AI chips on smartphones. By conducting all image classification and processing on the device itself, no images would need to be transmitted or stored externally. This approach could significantly alleviate users’ concerns about their images being stored or accessed by unauthorized parties. As AI technology continues to advance, incorporating on-device processing capabilities into our research methodology may not only increase user trust and comfort but also pave the way for a new
4.6 Discussion

In this section, we provide a summary of our findings and engage in a thorough discussion, exploring the implications and uncovering the potential opportunities highlighted by our results.

4.6.1. Summary of results

Our study investigated the potential of using in-the-wild smartphone images and deep learning models for detecting depression and predicting PHQ-8 scores, aiming to contribute to the development of user-centered and unobtrusive mental health assessment tools. The results of our analysis provided valuable insights into the characteristics of in-the-wild images, the performance of machine learning and deep learning models, and user acceptance of such approaches. The image characteristics analysis revealed that most images were captured from a low angle, indoors, and under well-lit conditions. These findings highlighted the participants’ natural behavior with their smartphones, emphasizing the importance of considering real-world HCI dynamics in designing mental health assessment tools.

Our predictive analysis demonstrates that a random forest model trained by manually selecting 3D landmark features obtains the best overall classification (balanced accuracy of 0.60, MCC of 0.14) and regression performance (MAE of 130.31). Interestingly, the EffNet deep learning model barely beat this score for classification task by 0.01. It correctly identified depressed and non-depressed participants with
a balanced accuracy of 0.61. Given additional high quality data, the deep learning models could improve over existing methods. To summarize, these scores are promising. They are even more noteworthy considering that the facial images were captured using a diverse range of smartphone devices – 87 different models from 9 distinct brands. As the camera quality of these devices varies significantly, it is important to note that the results may be influenced by factors such as image clarity and auto-focus capabilities. Despite these potential limitations, our findings support the ecological validity of the study and emphasize the potential of machine learning and deep learning methods in analyzing depression from facial images, even when captured in less-than-ideal conditions.

During post-hoc analysis, we gained several interesting insights. Firstly, our ablation study indicates that smaller domain-specific feature sets perform better in both our tasks. Specifically, we notice that 3D landmarks, gaze, and pose offer good performance across all metrics. By focusing on these features, researchers can potentially improve the overall performance of mental health assessment tools. Secondly, our explainability analysis revealed that larger values on the right side of the face have an impact on both depression detection and PHQ-8 score prediction. This finding suggests that people hold phones in a way that emphasizes the asymmetry of front-facing face images. Thirdly, our investigation into biases within machine learning models offers crucial insights for future research, particularly in terms of improving generalization and guiding data collection strategies. In terms of user acceptance, we found diverse responses regarding participants’ comfort levels with automated front-facing photo capture. While some participants were comfortable with the process, others felt uneasy due to concerns related to privacy, self-esteem, inappropriate situations, data security, and control over the images. These concerns highlight the need for careful consideration of ethical implications in designing and implementing studies.
involving photo bursts or similar data collection methods.

In conclusion, our research highlights the potential of using in-the-wild smartphone images, machine learning and deep learning models for depression analysis, offering a more objective, unobtrusive, and continuous approach to mental health assessment. By carefully considering the insights gained from our analysis and addressing the ethical implications, researchers and practitioners can work towards developing user-centered, effective, and ethically sound tools for mental health assessment and intervention.

4.6.2. Implications

The findings from our study hold significant implications for various stakeholders, including researchers, practitioners, and policymakers in the fields of mental health, digital health, human-computer interaction (HCI), and public health.

Our research highlights the potential of utilizing smartphone images and machine learning models as a supplementary method for mental health assessment. This innovative approach encourages the exploration of alternative ways to assess mental health that can complement traditional tools such as self-report questionnaires and clinical interviews. While our data was collected from participants who had major depressive disorder, the results pave the way for future research to investigate the broader applicability of these methods, potentially leading to a better understanding of depression and improved mental health support over time. Consequently, promoting timely access to appropriate interventions and support systems.

From an HCI perspective, our study underscores the importance of considering user acceptance when developing mental health assessment tools that utilize smartphone images and machine learning. Recently, there has been a growing interest among researchers to integrate user acceptance into the training phase of machine learning models, as proposed in studies like (187; 45). In a related observation, our
4.6 Discussion

Depression Detection Using Smartphone Images

feature importance analysis indicated that the right side of the face is more useful in depression detection. This phenomenon could be linked to the dominance of right-handed individuals, often resulting in partial face images that capture more of the right side. Various studies support the idea that handedness influences user interaction with smartphones and user experience (UX) (17; 228; 119; 175). Therefore, future research in HCI could benefit from focusing on developing tools that facilitate the capture of the entire face more effectively. For instance, the work by Nelavelli and Ploetz (228) explores adaptive app design tailored to the user’s handedness, which could be a promising direction for enhancing face image capture in smartphone applications. In summary, understanding users’ concerns and preferences is crucial for creating tools that are more likely to be adopted and used by those in need of support. This focus on user acceptance can inspire the HCI community to design mental health assessment tools that balance effectiveness, privacy, and user engagement, leading to the development of more accessible and inclusive digital mental health solutions.

In the broader context of public health, the study’s findings emphasize the importance of leveraging technology and innovative methods to address mental health challenges. As mental health disorders continue to impact individuals and communities worldwide, adopting novel approaches like the one presented in our study can contribute to more effective prevention strategies, early intervention, and resource allocation. This could ultimately lead to better mental health outcomes and overall well-being for individuals across various demographic and cultural contexts. In summary, the implications of our study extend well beyond the immediate findings, offering valuable insights for a range of stakeholders working at the intersection of mental health, digital health, and human-computer interaction. By considering user acceptance, exploring the potential of smartphone images for mental health assessment, and recognizing the broader public health context, our study contributes to
the development of more effective, user-friendly, and contextually appropriate mental health assessment tools with the potential to improve the lives of individuals affected by depression.

Section 4.7

Limitations

Our study while providing valuable insights into the use of in-the-wild smartphone images and deep learning models for depression detection, has some limitations that should be acknowledged. First, our study’s dataset may be limited in size and diversity, as it consists of a relatively small number of participants. Additionally, it is important to remember that our dataset is primarily composed of white females. Although our models currently show better performance for females in classification tasks and for non-females in regression tasks, expanding our dataset to include more diverse samples is necessary. By incorporating additional data that represents a broader spectrum of the population, we can ensure a more comprehensive representation. This expansion will not only enhance the robustness of our findings but also significantly improve the generalizability of our results across different demographic groups. Furthermore, the study relies on self-reported data, such as depression scores, which may be subject to biases, including social desirability and recall bias. Future research could be significantly enhanced by including more objective measures of mental health, such as clinical evaluations or physiological indicators. In our study, we adjusted each item’s score on the PHQ-8 from its original 0-3 range to a broader 0-100 scale. As mentioned earlier, the practice of re-scaling psychometric scales is not uncommon and has been applied to the PHQ in various past studies (196; 236; 127; 191). However, one limitation of adapting the PHQ-8 to a 0-100 scale is the potential for inconsistencies when correlating these scores with established levels of depression.
severity. To mitigate this, we proportionally scaled the original scores to derive our depression categorization, striving to preserve the original scoring system’s integrity. Additionally, our prediction models consider both the raw PHQ scores and the adjusted class scales, an approach that aims to balance detailed granularity with traditional scoring validity. It is also important to highlight that all participants in our study had received clinical diagnoses for MDD. However, we relied on self-reported data for tracking daily depression levels, which facilitated more consistent monitoring. Our study also focused exclusively on a clinically depressed cohort. Including healthy individuals in the dataset would have been beneficial for developing a more comprehensive and accurate prediction model. A randomized controlled trial (RCT) with healthy controls or incorporating a diverse cohort of individuals not experiencing depression could provide valuable insights into the differences between depressed and non-depressed individuals and improve the model’s ability to distinguish between them. Future research should consider expanding the dataset to include both depressed and healthy individuals, which can contribute to the development of more effective and precise mental health assessment tools.

Another limitation is that the study primarily focuses on the analysis of in-the-wild smartphone images and their relationship with depression. However, there may be other factors, such as social interactions, physical activity, and environmental context, that could provide additional insights into depression detection. Integrating these factors into future research may help to develop more holistic and accurate prediction models. Deep learning models, while powerful and effective, can often be considered as "black-box" models with limited interpretability. This may make it difficult to understand the specific features or patterns that the model has identified as being related to depression. Future research could explore the use of more interpretable models or techniques to provide insights into the underlying mechanisms linking visual
4.8 Conclusion

Depression Detection Using Smartphone Images

Cues and depression. Lastly, the use of in-the-wild smartphone images for mental health assessment raises ethical and privacy concerns, which need to be carefully considered when designing and implementing such tools. Ensuring user consent, data security, and transparency in the use of personal data is crucial for maintaining trust and fostering the adoption of these tools. Addressing these limitations in future research can help to further advance our understanding of the relationship between smartphone images, deep learning models, and depression detection, contributing to the development of more effective, user-centered, and ethically sound mental health assessment tools.

Section 4.8

Conclusion and Future Work

Through this study, we have demonstrated the potential of using in-the-wild smartphone images and machine learning to detect depression, offering valuable insights for mental health assessment, HCI and digital health. With this, we aim to pave the way for more effective and user-centered mental health assessment tools. Addressing the limitations of our study and building upon its findings, future research can contribute to the development of more robust, accurate, and ethically sound mental health assessment tools that have the potential to improve the lives of individuals affected by depression.

When we embarked on designing our MoodCapture study to investigate whether high-resolution face capture from phones could assess mood, we were acutely aware of the ethical issues surrounding our research and the potential privacy concerns of a population that included individuals diagnosed with depression. As discussed in the section on Ethical Considerations and User Acceptance, our study was meticulously designed to safeguard user privacy throughout, and we sought their evaluations of the
MoodCapture app post-study. This invaluable feedback forms the foundation for future work in image-based mood detection which we believe is a promising technology. One direction we plan to pursue as our next step involves utilizing on-phone AI chips that are now available on top-end smartphones to run deep learning models directly on the device, ensuring that images never leave the phone. Additionally, we intend to explore the combination of this on-device prediction approach with federated deep learning, where models are trained without sharing raw data across a network in a central entity such as a server or cloud. This approach could effectively address security concerns associated with centralized data collection and the privacy issues our participants raised during the acceptance study. Finally, we recognize that the performance of the models we considered for face-based depression detection, particularly deep learning models, would benefit significantly from a larger face dataset. In the MoodCapture study, we collected over 125,000 images from 177 individuals living with depression over a period of 90 days, representing a well-sized dataset to demonstrate the potential of this idea. If future face-based depression studies have access to larger pools of naturalistic images (e.g., VGGFace2, which contains over 3 million face images) collected in the wild, we anticipate that the accuracy and capabilities of the models would see significant improvement.

The research discussed in this chapter was supported by the National Institute of Mental Health (NIMH) under award number R01MH123482-01. We acknowledge that the content of this chapter is solely our responsibility and does not necessarily reflect the views of the NIMH. We also clarify that the funding body had no involvement in the study’s design, data collection, analysis, interpretation, or chapter preparation.
Chapter 5

From User Surveys to Telemetry-Driven Agents: Exploring the Potential of Personalized Productivity Solutions

Section 5.1 Introduction

In today’s dynamic work environment, productivity plays a pivotal role in determining professional success. Organizations and individuals continuously strive to enhance their efficiency, improve time management, and optimize their work processes. The rapid advancements in artificial intelligence (AI), natural language processing, and machine learning technologies have introduced new opportunities to develop tools that can assist in boosting productivity. Several studies have explored the potential of AI-based tools in various professional contexts, such as project management (306; 77), team collaboration (26), and decision-making (143). These studies have demonstrated
the potential of AI to not only automate routine tasks but also to provide intelligent insights and recommendations that can help users optimize their work processes and achieve their goals more efficiently. To develop a truly effective productivity agent, it is crucial to understand user preferences and the factors that influence productivity. This paper presents a comprehensive two-phase study aimed at investigating user preferences for productivity agents and devising personalized solutions that cater to these preferences. In the first phase, we conducted a user survey to identify the ideal persona for a productivity agent by examining people’s expectations and preferences concerning its features and behaviors, exploring various aspects such as productivity, communication style, agent approach, personality traits, personalization, and privacy. Our findings revealed users’ specific needs and areas where a personalized productivity agent could offer the most significant benefits. We also found that striking a balance between personalization and privacy is essential, as well as the need for adaptability in tone and approach depending on the situation and user preferences.

In the second phase, we developed an AI-based productivity agent that leverages insights from the initial user survey and uses the state-of-the-art GPT-4 model (246). GPT-4 is a powerful Large Language Model (LLM) (227) capable of generating human-like text and understanding context, making it a reasonable choice for creating a personalized productivity agent. To facilitate personalization, information workers participating in our study provided their workplace telemetry data, such as email and meeting habits, to the productivity agent through a custom application that we developed. The agent used the shared high-level telemetry data to tailor its recommendations and insights based on each user’s unique work habits and patterns, providing personalized assistance for a more effective and customized user experience. In a within-subjects study, we then evaluated the AI agent’s performance and compared it with alternative methods of telemetry data presentation, such as dash-
board visualization and personalized narratives. This analysis allowed us to better understand user preferences and identify areas for improvement and refinement.

Our objective in this comprehensive study was to delve into the nuanced requirements and preferences of users regarding AI-driven productivity tools. To structure our investigation, we formulated specific research questions for each phase of the study. These questions guided our research methodology and helped us focus on key areas of user interest and technology application:

- **(RQ1):** What are the ideal characteristics and capabilities desired by users in a productivity agent? This question aimed to uncover the specific features, traits, and functionalities users seek in a productivity-enhancing AI tool, considering aspects like personalization, communication style, and privacy concerns.

- **(RQ2):** How does the AI productivity agent based on findings from RQ1 compare with traditional methods of data presentation (dashboard visualization and narrative summaries)? This question seeks to evaluate the effectiveness of the AI productivity agent, designed in line with RQ1 findings, against more conventional data presentation methods. It focuses on user preference, engagement, and the perceived utility of each approach in enhancing workplace productivity.

- **(RQ3):** How effective is the AI productivity agent in understanding and addressing users’ productivity challenges based on their telemetry data? This question delves into the AI productivity agent’s performance, particularly focusing on its capability to analyze and respond to individual user productivity needs using the gathered telemetry data.

- **(RQ4):** What are the perceived strengths, weaknesses, and potential areas for improvement in the AI productivity agent, visual dashboard, and AI-based personalized narrative? This question seeks to gather insights on the comparative
advantages and limitations of the three data presentation methods, as well as identify opportunities for enhancing their overall performance and user experience.

In addressing the research questions discussed above, we make the following contributions:

- We conducted a comprehensive user survey with N=363 information workers that revealed users’ specific needs and preferences for productivity agents. For example, we found that the majority of the participants preferred the agent to provide proactive suggestions to enhance productivity and that they were comfortable sharing high-level telemetry data for enhancing the response provided by the productivity agent.

- We developed a custom application that facilitated user telemetry data sharing, enabling the AI-based productivity agent to offer personalized assistance. The application would scrape telemetry data off of web pages. We utilize the GPT-4 model for extracting information from the web page as well to create the personalized AI productivity agent.

- We conducted a thorough evaluation of the AI agent’s performance with N=40 participants, highlighting its ability to provide personalized insights and recommendations based on user telemetry data. In addition, we compared the AI agent with alternative methods of telemetry data presentation, indicating user preferences for a combination of dashboard visualization and interactive chat agents.

As we continue to witness rapid advancements in AI, natural language processing, and machine learning technologies, there is a growing potential for further enhancing and refining productivity agents. The incorporation of LLMs in the development
of our AI-based productivity agent demonstrates the power of language models in creating personalized, context-aware, and potentially efficient productivity-enhancing tools. Our findings underscore the importance of a user-centric approach to the design and development of AI-based productivity agents and highlight the potential of incorporating design principles, such as semantic zooming (291) and progressive disclosure (239), to enhance their effectiveness and user experience. By building on the insights gleaned from this study, future research can continue to refine and optimize productivity-enhancing tools and solutions, ultimately leading to improved efficiency and user-centric experiences for information workers. This paper makes meaningful contributions to the field of HCI and AI-assisted productivity, and we hope these can meaningfully guide practitioners in the development of effective, user-centric solutions in this domain.

Section 5.2

Related Work

The landscape of productivity tools has undergone a significant evolution in recent years. The field has progressed from rudimentary task managers and calendar applications to more sophisticated AI-driven virtual assistants, such as Microsoft’s Cortana, Apple’s Siri, and Google’s Assistant. While numerous studies have emphasized the effectiveness of these general-purpose virtual assistants in enhancing user efficiency and satisfaction across various contexts and domains (80; 160; 52), there remains a gap in exploring AI-driven productivity agents tailored specifically to individual needs. This gap becomes even more pronounced when considering the unique demands of information workers, who face challenges that often diverge from those of the broader workforce or consumers.

Information workers, who primarily engage in data-centric tasks, often require
more specialized tools to address their unique professional demands. While general-purpose virtual assistants offer some degree of support, they may not fully cater to the distinct requirements of these professionals such as scheduling meetings, summarizing research reviews, and preparing presentations or reports. In recent years, customized solutions—particularly conversational agents—have emerged as a promising alternative due to their adaptability to individual professional needs (88; 99). However, the majority of existing research on conversational or chat agents is centered on customer service (98; 75) or marketing and sales (322; 326). Other common areas of study for agents include education (131) and health (42; 32). Despite this focus, some progress has been made in researching solutions for information workers. For example, various conversational agent-based approaches have been proposed to assist with scheduling and managing tasks (317; 115), managing distractions (319; 199), coordinating meetings (74), work reflections (163) and even augmenting group decision-making (289). Notably, the work by Grover et al. (125) demonstrated the design of a chatbot-like agent, complemented by an emotionally expressive video avatar, that aimed to enhance productivity. Their productivity agent assisted N=40 users in scheduling focused tasks, monitoring distractions, and reflecting on their daily mood and goals through a standalone application. Authors reported that the agents led to an increase in scheduled time for focused tasks and that the users felt more satisfied and productive with the agents. Similarly, Kimani et al. (159) introduced Amber, a conversational agent designed to support information workers across a diverse range of tasks, such as scheduling, prioritizing tasks, task switching, providing break reminders, managing social media distractions, and reflecting on daily accomplishments. The researchers conducted a field study with 24 participants over six days and found promising results for the potential use of conversational agents in enhancing workplace productivity and well-being. A critical aspect of designing these
agents is understanding user preferences. For example, Kimani et al. (159) conducted an online survey with N=70 participants to investigate when information workers might engage with a conversational agent. The authors found that users needed help with reminders, scheduling tasks, and managing distractions. Another study by Ahire et al. (6) explored the expectations of N=14 knowledge workers from such agents, with findings highlighting the importance of features related to scheduling, distraction management, and well-being. Khaokaew et al. (153) conducted a user study of N=40 workers over four weeks to identify user needs for not just conversational agents but for all kinds of digital assistants (DA). The authors report that the participants imagined a DA that supported them by managing their (1) time, (2) tasks, and (3) information. While these insights provide valuable information, there is still a pressing need for a more comprehensive exploration of the specific requirements of information workers. Previous studies have broadly addressed productivity tools without honing in on the unique challenges and needs of information workers. Our research aims to fill this gap by focusing specifically on how these professionals interact with and benefit from AI-driven productivity assistants. Note, in this study, ‘information workers’ refer to professionals whose primary role involves acquiring, manipulating, and generating information (173).

Recent advancements in natural language processing, particularly the emergence of advanced language models such as GPT-3 and GPT-4, have significantly expanded the potential of AI-driven conversational agents. However, creating agents tailored to individual users’ work habits and patterns remains relatively uncharted territory. Although earlier applications of conversational agents in AI productivity tools, as developed by Grover et al. (125) and Kimani et al. (159), utilized telemetry or sensing data for customization, it has yet to be thoroughly explored, particularly with the latest large language model-based chat agents that offer a more flexible and power-
ful alternative to their predecessors due to their semantic understanding. Telemetry
data, which provides valuable insights into user behaviors and preferences, is essential
for achieving meaningful personalization. Despite its potential to enhance AI-driven
productivity agents, the implementation of telemetry data in this context has been
limited. Our research aims to bridge these gaps by conducting a comprehensive two-
phase study that examines user preferences and incorporates telemetry data for per-
sonalization in AI-driven productivity agents. Our approach goes beyond the existing
literature by offering an in-depth analysis of both user preferences and the affordances
of telemetry data. We delve deeper into its potential by integrating it with advanced
AI capabilities, offering a more nuanced and personalized user experience. The use
of large language models (LLMs) brings enhanced capabilities in understanding and
processing complex user data, providing more accurate and contextually relevant pro-
ductivity insights. Thus, our study’s innovation lies in the use of cutting-edge LLMs
and how we leverage these models to interpret telemetry data, creating a personalized
and dynamic productivity tool tailored to individual user needs. We have the position
that our findings can pave the way for future research and contribute to the develop-
ment of more sophisticated productivity tools tailored to the distinct challenges and
needs of information workers.

Section 5.3 Methodology

In this section, we present an overview of our study methodology, encompassing its
design, data collection, and participant demographics. Our research methodology is
divided into two distinct phases (as shown in Figure 5.1). The first phase involves con-
ducting a user survey to gain insights into people’s preferences regarding productivity
agents. Utilizing the information gathered from the survey, the second phase focuses
5.3 Methodology

**Personalized Productivity Solutions**

Figure 5.1: Two-Phase Study: In Phase 1, we explore desired productivity agent traits, while in Phase 2, we design and compare it with a visual dashboard and personalized narrative.

on designing a personalized productivity agent that caters to these user preferences. This two-phase approach ensures that our productivity agent designs are informed by users’ needs and preferences, ultimately leading to more effective and user-centric solutions. All of the participants we enrolled are information workers working at a large multinational technology company based in the United States. The study was approved by the study institution’s Internal Review Board (IRB).

5.3.1. Understanding User Preferences for Productivity Agents

The primary objective of our first survey was to identify the ideal persona for a productivity agent by examining people’s expectations and preferences concerning its features and behaviors. We conducted a user survey with N=363 information workers, using this as an initial step to gather insights into their desired characteristics and capabilities for a productivity agent. The survey delved into various design aspects, including the agent’s empathy level, proactive assistance, communication style, and adaptability to different situations. The survey we designed was carefully crafted based on prior work (125; 159) and extensive interactions with experts in the field (i.e., AI professionals, organizational psychologists with expertise in workplace productivity and technology adoption, productivity engineers and expert HCI researchers). Drawing upon their knowledge and experience, we aimed to create a
5.3 Methodology

Personalized Productivity Solutions

comprehensive and insightful questionnaire that would effectively capture users’ preferences and concerns in relation to a productivity agent. Refer to Appendix B.1 for a comprehensive list of questions posed to participants. We distributed the survey randomly to employees within the company and received a total of 363 responses. The average duration participants spent on the survey was 17 minutes. Participation was anonymous and there was no compensation for this phase of the study.

Demographics: Appendix B.6 presents the demographic information of our participants. The majority of the participants identified as male (67.2%, N=244). In terms of age group, most participants were in the 36-45 (27.5%, N=100), 46-55 (26.7%, N=97), and 26-35 (26.4%, N=96) age ranges. The majority of participants worked in software development/engineering roles (46.3%, N=168), followed by product management (19.6%, N=71).

5.3.2. Productivity Solutions Design and Evaluation

In the following phase of our study, we concentrated on developing an AI productivity agent specifically aimed at enhancing productivity. The design of the productivity agent was based on the insights obtained from the initial user survey, ensuring that the agent addresses the preferences and expectations of the participants. To further investigate the effectiveness of the proposed agent, we also asked participants to interact with two alternative methods of presenting telemetry data: a traditional dashboard visualization and an AI-based personalized narrative.

To personalize the experience to each individual, we decided to use the high-level telemetry data on participants’ meetings, communication, collaboration network, and other work habits which are provided through Microsoft’s Viva Insights\(^1\). Among the different platforms, we selected Viva Insights as it is widely used in many companies

\(^1\)https://learn.microsoft.com/en-us/viva/insights/Use/Metric-definitions
5.3 Methodology

Personalized Productivity Solutions

Figure 5.2: Enhanced Telemetry Data Scraping: Figure (a) displays the user’s initial view as our advanced scraping tool navigates through various Viva Insights pages to capture their source code. Once the source code is analyzed and the necessary data extracted, users will be presented with the final screen depicted in Figure (b). This concluding screen showcases the retrieved information, allowing users to review and modify the telemetry data prior to submission.

and computes a wide variety of relevant features in the context of productivity. To extract the information, we built a custom dashboard scraper (see Figure 5.2) that allowed participants to automatically extract their Viva Insights data, review and modify it, and ultimately share it with the researchers. The extracted data was then used to create the three modes of presentation tailored to each participant’s work patterns and habits. After participants willingly shared their telemetry data with us, they engaged with the three modes: viewing their data on a dashboard, reading a
5.3 Methodology Personalized Productivity Solutions

personalized narrative, and interacting with AI productivity agent that had access
to their telemetry data. They then assessed their experiences based on the ability to
gain productivity insights or identify opportunities for productivity improvement by
answering our follow-up survey. Note, the AI productivity agent primarily functioned
as a ‘productivity guide,’ offering recommendations and advice to improve users’ pro-
ductivity. This included suggesting strategies for time management, prioritizing tasks,
and optimizing work routines. However, it is important to note that the agent did
not directly perform tasks such as meeting summarization or documentation writing.
While it provided guidance in these areas, the execution of tasks remained the user’s
responsibility. This distinction is crucial for understanding the agent’s role in our
study. In total, we had N=40 participants complete the second phase of our study.
To diversify our sample, we reached out to individuals who had not participated in the
first phase of the study. The average time for study completion was 30.15 minutes,
and participants received a $25 gift card as compensation. Our study exclusively
makes use of Open AI’s (247) GPT-4 API.

Table 5.1: Telemetry: List of telemetry data collected from the participants using our
scraper tool.

<table>
<thead>
<tr>
<th>Category</th>
<th>Telemetry Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting</td>
<td>total number of meetings organized and attended</td>
</tr>
<tr>
<td></td>
<td>percentages of meetings organized that were one hour or shorter</td>
</tr>
<tr>
<td></td>
<td>meeting invitations sent with more than 24 hours notice before the scheduled start time (i.e. advanced notice)</td>
</tr>
<tr>
<td></td>
<td>meetings organized that included a Teams link for remote attendees</td>
</tr>
<tr>
<td></td>
<td>meetings organized or accepted during working hours</td>
</tr>
<tr>
<td></td>
<td>online meetings (on Microsoft Teams) that ended within one minute of the scheduled end time</td>
</tr>
<tr>
<td></td>
<td>meetings organized or accepted that had a response rate over 50%</td>
</tr>
<tr>
<td></td>
<td>online meetings (on Microsoft Teams) joined within five minutes of the scheduled start time</td>
</tr>
<tr>
<td></td>
<td>meetings that did not overlap with other meetings on the calendar</td>
</tr>
<tr>
<td></td>
<td>meetings invited to and either accepted or declined (i.e., not having Tentative status)</td>
</tr>
<tr>
<td></td>
<td>meetings where the user did not read or send emails or chats (i.e., no multitasking)</td>
</tr>
<tr>
<td>Communication</td>
<td>total number of emails (sent and read)</td>
</tr>
<tr>
<td></td>
<td>hourly breakdown of emails sent, emails read, and chats/calls made</td>
</tr>
<tr>
<td>Collaboration</td>
<td>number of collaborators</td>
</tr>
<tr>
<td></td>
<td>collaboration time within and outside working hours</td>
</tr>
<tr>
<td>Workplace Wellbeing</td>
<td>total duration of focus time (i.e., uninterrupted work time) kept along with reserved focus time for next week</td>
</tr>
<tr>
<td></td>
<td>number of days with disruptions (i.e. emails, chats, and meetings) after work hours</td>
</tr>
</tbody>
</table>

Rationale for Comparative Analysis: The transition from Phase 1 to Phase 2 of
our study was guided by the need to explore the practical application of user preferences identified in the initial survey. While Phase 1 established the preferred characteristics and functionalities of a productivity agent, Phase 2 aimed to contextualize these preferences within actual work settings. The introduction of a dashboard and narrative alongside the AI agent was a strategic decision to compare different modes of data presentation. This comparison was crucial to understanding how different formats could align with or diverge from the identified user preferences. The dashboard was considered for its traditional, structured approach to data visualization, while the narrative format was explored for its potential to provide a more engaging and story-like presentation of data. By comparing these with the AI agent, we sought to evaluate the effectiveness, user engagement, and preference alignment of each method, thereby offering a comprehensive view of how different data presentation strategies can cater to the diverse needs and preferences of information workers. This comparative analysis not only validates the effectiveness of our AI agent but also contributes to a broader understanding of user-centric data presentation in productivity tools.

**Telemetry Data Collection:** Our scraper tool employs Selenium (288), a testing library, to open up a web browser and navigate the Viva Insights pages, capturing the HTML source of each page. Since these pages are dynamically generated, we opted for a more error-resilient approach by utilizing GPT-4 to extract the desired information, rather than relying on traditional HTML tag-based scraping. Once we obtained the HTML source, we removed unnecessary HTML tags and attributes and then sent the streamlined source to GPT-4 with a specific prompt to help identify the information we sought. This method proved to be effective in facilitating the extraction of relevant data from the Viva Insights pages. Please refer to Appendix B.3 and Appendix B.4 for the comprehensive list of system prompts and user prompts.
we provided to GPT-4 for extracting and transforming the desired information. The telemetry information obtained from Viva Insights can be found in Table 5.1. It is important to note that Viva Insights calculates these insights based on the data from the prior month.

**Designing Productivity Solutions:** As mentioned earlier, we designed three productivity solutions: an AI productivity agent (also referred to as an interactive chat agent), a personalized narrative (generated via GPT-4), and a traditional visual dashboard. We began by integrating telemetry data directly into the AI agent’s system prompt, a crucial step that empowered the agent with the ability to analyze and respond based on individual user behaviors and patterns. This data-driven approach ensured that each participant received feedback and suggestions that reflected their unique work habits. To ensure the agent’s feedback and recommendations resonated with users, we drew from the insights gathered during the first phase of our study. The preferences and expectations participants shared during this phase became the foundation for designing the system prompt of the AI agent. Parallel to the development of the AI agent, we embarked on creating an AI-based personalized narrative, an alternative means of communicating productivity insights. This narrative was also crafted to align with the specific preferences participants had voiced. Beyond just catering to these preferences, we introduced a structured format to this narrative, ensuring as consistent delivery of insights as we could across participants. The narrative format presented information in a structured, story-like manner, providing users with an overview of their productivity patterns and suggesting improvements. In contrast, the interactive agent engaged users in a conversational interface, offering personalized guidance based on user inputs and queries. This interactive nature allowed for a more tailored and dynamic exchange, though it was limited to offering
5.3 Methodology

Personalized Productivity Solutions

(a) Dashboard visualization

(b) Personalized narrative

(c) AI productivity agent

Figure 5.3: Designing Productivity Solutions: Figure (a) displays the dashboard presentation, (b) shows the AI-based personalized narrative, and (c) the AI productivity agent.
advice rather than executing tasks directly. To perfect the prompts for both the AI productivity agent and the personalized narrative, we engaged in multiple cycles of testing and refinement. Each iteration was an opportunity to align more closely with user needs and preferences. Refer to Appendix B.5 for the full prompts used for the productivity agent and personalized narrative. For the development of the dashboard visualization, we recognized that many participants had established familiarity with Viva Insights, so we aimed to make our dashboard’s visual presentations echo those of Viva Insights. We meticulously designed a range of visual representations, from bar charts that provided at-a-glance insights to pie charts that broke down data into comprehensible segments. Furthermore, we chose to present certain telemetry data numerically, ensuring users had access to precise figures. Figure 5.3 presents the AI productivity agent along with the other two presentation modes.

**Evaluation Procedure**: The evaluation process began with participants using our data scraper tool to collect their Viva Insights information, which was then shared with us upon their review. Each presentation mode incorporated this high-level telemetry data from individual participants, allowing for the provision of insights (in the case of the dashboard) and tailored assistance and recommendations (in the case of both the personalized narrative and the AI productivity agent) to improve productivity. Consequently, participants interacted with the three presentation modes, seeking guidance on enhancing their productivity based on their unique work patterns and habits. The study flow is outlined below:

(a) Participants used the scraper tool to gather and share the telemetry data with the research team, identified by a unique code assigned to each participant. The tool submitted three types of files to our storage: a) the raw telemetry data, b) the telemetry data transformed to JSON format using GPT-4, and c) the
5.3 Methodology

personalized narrative generated using GPT-4.

(b) We provided each participant with a Qualtrics survey link with their unique code embedded in it. This survey included basic demographic questions and links to access the three presentation modes.

(c) To mitigate potential ordering effects, participants engaged with the three presentation modes in a randomized sequence and evaluated each mode. The survey displayed the links to the websites featuring the dashboard, interactive chat agent, or personalized narrative in a random order. When participants clicked on the links, the web pages opened with the participant’s unique code embedded as part of the query parameter.

(d) Since the study aimed to assess the duration of engagement with each mode, participants were allowed unlimited time for exploration. However, each presentation mode instructed participants to familiarize themselves with the presentations’ capabilities and insights, as questions about their functionality might be asked. This ensured full engagement with the chosen presentation mode. As the presentation mode had access to the participant’s unique code as part of its query parameter, it could refer to the participant’s data in our storage. For the dashboard, this meant reading the JSON-transformed telemetry data of the participant and using it to populate the visualizations. For the personalized narrative, it would load the relevant pre-generated narrative, ensuring that participants did not need to wait for their personalized narrative to be generated once they opened the website. For the interactive chat agent, this meant reading the telemetry data and providing it as part of the system prompt to the chat agent. To complete their interaction, participants clicked the "Finish" button available on top of each page, and the duration of their engagement was
recorded for further analysis. After clicking "Finish," they were presented with a passcode that they entered into their survey to proceed to the next presentation mode. Access to the next mode was granted only upon submission of the correct passcode. After entering the correct passcode, participants answered questions related to the current presentation mode before advancing to the next one. We maintained consistent page design for each of the three presentation modes, i.e., the buttons looked the same and were placed in the same position, and the colors, spacing, and other visual elements remained the same.

(e) After completing all three modes and evaluating them through a series of survey questions, participants ranked the modes according to preference and provided any pros, cons, or suggestions. See Appendix B.2 for the list of questions asked during the evaluation.

**Demographics:** Appendix B.7 presents the demographic information of our participants. Most of the participants identified as men (60.0%, N=24), while 40.0% (N=16) identified as women. In terms of age group, the largest proportion of participants was in the 26-35 (30.0%, N=12) and 46-55 (32.5%, N=13) age ranges, followed by 36-45 (17.5%, N=7), 56-65 (12.5%, N=5), and 18-25 (7.5%, N=3) age ranges. Regarding job roles, the majority of participants worked in product management (52.5%, N=21), with software development/engineering being the second most common role (10.0%, N=4). Other roles included administrative/operations (7.5%, N=3), data science/analytics (7.5%, N=3), customer support (5.0%, N=2), IT/infrastructure (2.5%, N=1), and other roles (15.0%, N=6).
In this section, we provide an overview of the results obtained from both phases of our study: the user surveys and the design and evaluation of the productivity agent.

5.4.1. Exploring User Preferences for Productivity Agents

In the first phase of the study, we conducted a comprehensive user survey to gather insights into participants’ preferences and needs concerning a personalized productivity agent. The survey aimed to explore various aspects of productivity, communication style, agent’s approach, personality traits, personalization, and privacy. We divide the results into themes based on the main topics of the survey questions to better understand the participants’ perspectives and tailor the productivity agent we design in the second phase to meet their specific requirements effectively. Please refer to Appendix B.8 for the full list of responses.

**Productivity Perspectives.** We focused on understanding participants’ perspectives on productivity, the usefulness of a productivity agent, the agent’s personality traits, and the areas in which they would like assistance. The large majority of participants (95.6%) reported that making progress or completing scheduled tasks would make them feel productive at the end of the day. Solving novel, unexpected tasks (76.6%) and helping a colleague accomplish a task (72.2%) were also significant contributors to their feelings of productivity. Participants could select multiple options for this question.

Out of all the respondents, 280 (77.1%) expressed that a productivity agent would be beneficial to them. Consequently, we presented the remaining survey questions only to these participants, who expressed a need for a productivity agent. Among them, 171 (61%) believed that the personality of the productivity agent would have a
5.4 Results

Personalized Productivity Solutions

(a) Desired areas of help with productivity agent

Figure 5.4: Productivity Areas and Language: Figure (a) illustrates the area that the participants struggle with or would like an AI productivity agent to help them with. Figure (b) shows the preference of participants with regard to the agent’s language or tone.

Note that in our study’s second phase, we consciously selected productivity-enhancing features such as time management, meeting scheduling, and work summaries for our AI agent’s focus. This decision was based on substantial user interest and the suitability for immediate application and testing within diverse work settings. These areas, while not the highest in demand according to our Phase 1 findings, presented a unique opportunity to evaluate the real-world efficacy of AI-driven productivity advice. Furthermore, the inclusion of these features enabled a comparative moderate-to-significant impact on its effectiveness as a productivity tool. As shown in Figure 5.4a, when asked about the areas in which participants would like an AI agent to provide assistance, the top five categories were: 191 meeting summaries (68.2%), 176 documentation writing (62.9%), 159 start-of-day briefs (56.8%), 157 work summaries (56.1%), and 150 information retrieval (53.6%). These findings provide valuable insights into the specific tasks and areas where a personalized productivity agent can offer the most significant benefits to its users.
5.4 Results

Personality Productivity Solutions

analysis with traditional data presentation methods like dashboards and narrative summaries. This comparison was instrumental in assessing user experience and preference, providing a clearer understanding of the potential and limitations of AI agents in enhancing productivity. The insights gained here are invaluable for directing the evolution of AI productivity tools, ensuring they align more closely with user needs and preferences in subsequent developments.

Communication Style. We explored participants’ preferences regarding the tone and style of communication, as well as their desired level of rapport-building with the productivity agent. A total of 99 participants (35.4%) expressed their preference for a communication tone that combines elements of both casual and formal styles, maintaining a friendly demeanor while still conveying professionalism. Additionally, 95 participants (33.9%) mentioned that their preferences for communication tone varied depending on the specific task or situation. As shown in Figure 5.4b, when it came to rapport-building, 71 participants (25.4%) responded that a professional and respectful language or tone would make them feel most comfortable opening up to the productivity agent about concerns related to productivity. Similarly, 69 participants (24.6%) preferred a direct and straightforward tone.

Regarding the style of messages, 99 participants (35.4%) responded that they would prefer direct and to-the-point messages from the productivity agent, while 86 (30.7%) said that they would prefer a mixture of chatty and conversational as well as direct and to-the-point. These findings highlight the importance of incorporating a balance between casual and formal communication styles, as well as the need for adaptability in tone, depending on the situation and user preferences.

Approach & Personality Traits. We investigated the participants’ preferences regarding the productivity agent’s approach to task assistance, desired personality traits, and preferred social roles.
5.4 Results

5.4.1 Personalized Productivity Solutions

(a) Desired personality traits

(b) Preferred social roles

Figure 5.5: Personality Traits and Social Roles: Figure (a) depicts the different personality traits participants prefer in the productivity agent. Figure (b) shows the social role desired by the participants for the agent.

Most of the participants, 169 (60.4%), expressed their preference for the agent to provide proactive suggestions to enhance productivity. In contrast, 77 participants (27.5%) wanted the agent to provide assistance and suggestions only when explicitly asked for help. The agent’s preferred personality traits are depicted in Figure 5.5a; the most common preferences were: knowledgeable (88.9%), trustworthy (75.0%), and transparent (68.2%). Similarly, Figure 5.5b represents the social role that people desired for the agent. The top choices were that of an assistant (76.4%) and expert (56.1%). The supervisor role was the least preferred, with only 4.6% of participants opting for it. These findings emphasize the importance of designing a productivity agent with a proactive approach and equipping it with the desired personality traits, such as knowledgeability, trustworthiness, and transparency, to better cater to users’ preferences.

**Personalization & Privacy.** We explored participants’ preferences regarding personalized recommendations and privacy concerns when using a productivity agent.

A significant majority of participants, 209 (74.6%), expressed that they would
prefer a moderate to high level of personalization in the productivity agent. This enables the agent to provide highly personalized advice tailored to their specific needs, considering various aspects of their work habits and preferences. In terms of privacy, 152 participants (54.3%) felt either somewhat comfortable or completely comfortable sharing their personal information with the productivity agent. However, 58 participants (20.7%) expressed discomfort in sharing their personal information. Interestingly, a majority of participants, 205 (73.2%), were either somewhat comfortable or completely comfortable sharing high-level telemetry data for enhancing the quality and guidance provided by the productivity agent. These findings highlight the importance of striking a balance between personalization and privacy to cater to users’ preferences and concerns.

**Negative traits to avoid.** We asked participants an open-ended question – “In your opinion, what are some potential negative personality traits that should be avoided in a productivity agent to ensure a positive, effective, and supportive user experience?” The responses emphasized the importance of avoiding traits such as nagging, aggressiveness, pushiness, verbosity, overly casual or friendly demeanor, judgmental or patronizing tones, and condescension. Participants also warned against providing false or exaggerated praise, taking on a “taskmaster” or micromanager role, and presenting inaccurate information. Other concerns included anthropomorphism, avoiding complex language or exclusionary phrasing, and being too intrusive. Participants generally preferred an agent that acts as a helpful assistant, rather than a replacement for a counselor or a parent.

5.4.2. Design of the AI Productivity Agent Informed by User Preferences

We now present how the design of the AI productivity agent was crafted based on the insights obtained from the initial user survey conducted in Phase 1 (Section 5.4.1). This survey comprehensively explored users’ preferences and expectations in
5.4 Results

PERSONALIZED PRODUCTIVITY SOLUTIONS

Table 5.2: Design of the AI Productivity Agent Based on Phase 1 Survey Findings.

<table>
<thead>
<tr>
<th>Theme</th>
<th>User Preference (Phase 1)</th>
<th>Integration in Productivity Agent (Phase 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productivity Perspectives</td>
<td>Participants felt productive when making progress on scheduled tasks, solving unexpected tasks, and helping colleagues</td>
<td>Agent designed to offer strategies for task management, prioritizing, and collaborative assistance. It is developed to help users achieve an optimal work-life balance and improve their productivity.</td>
</tr>
<tr>
<td>Communication Style</td>
<td>Preference for a tone that balances casual and formal styles; adaptability in tone based on task and situation; direct and straightforward communication rather than chatty</td>
<td>Agent uses a professional yet friendly tone, adaptable to the context and user’s current task. It is designed to communicate concisely without being overly chatty or verbose.</td>
</tr>
<tr>
<td>Approach</td>
<td>Desire for proactive suggestions and assistance to enhance productivity</td>
<td>Agent designed to provide proactive productivity suggestions without being too intrusive and pushy.</td>
</tr>
<tr>
<td>Personality Traits</td>
<td>Valued traits included being knowledgeable, trustworthy, and transparent.</td>
<td>Agent developed to demonstrate these traits consistently across interactions.</td>
</tr>
<tr>
<td>Personalization</td>
<td>Preference for a moderate to high level of personalization, balancing with privacy concern</td>
<td>Agent uses high-level telemetry data to personalize suggestions while ensuring user privacy.</td>
</tr>
<tr>
<td>Negative Traits to Avoid</td>
<td>Avoidance of traits like being nagging, aggressive, pushy, overly casual, judgmental, or patronizing.</td>
<td>Agent prompted to avoid these negative traits, focusing on supportive and informative interactions.</td>
</tr>
</tbody>
</table>

several key areas: productivity perspectives, communication style, approach, personality traits, personalization, and desired and undesired traits in an AI agent. A summary of these findings and their corresponding implementations in the AI agent’s design is presented in Table 5.2.

The table illustrates how each identified theme from the survey was translated into specific features of the AI productivity agent. For instance, user preferences for a balanced communication style influenced the agent’s development to ensure it communicates in a professional yet approachable manner. Similarly, the agent’s approach
and personality traits were tailored to align with users’ desires for proactive assistance and valued characteristics like knowledgeability and trustworthiness. Additionally, personalization emerged as a crucial aspect, with users expressing a need for tailored recommendations balanced against privacy concerns. Accordingly, the agent leverages high-level telemetry data to offer personalized suggestions, carefully designed to respect user privacy. This careful alignment of the AI productivity agent’s features with user preferences underscores our commitment to developing a user-centric tool. It highlights the importance of basing design decisions on thorough user research, ensuring that the developed AI agent is not only technologically advanced but also resonates with the actual needs and preferences of its intended users.

5.4.3. Productivity Solution Showdown: AI Agent vs. Dashboard vs. Personalized Narrative

In this section, we present the findings from the second phase of our study. Specifically, we explore the interaction duration, participants’ overall experiences, anticipated usage frequency, and preference rankings for the three presentation methods: the AI productivity agent, visual dashboard, and AI-based personalized narrative.

Participants spent an average of 60.58 seconds (with a standard deviation of 38.98 seconds) interacting with the dashboard. In comparison, they spent considerably more time on the AI productivity agent, with an average duration of 315.48 seconds (standard deviation: 272.23 seconds). The personalized narrative had a similar interaction time to the dashboard, with an average of 65.21 seconds (standard deviation: 55.39 seconds). After interacting with each method, participants were asked to rate their overall experience on a six-point Likert scale ranging from Very Poor to Excellent. As shown in the stacked bar graph in Figure 5.6a, 26.2% of participants rated the Dashboard as excellent, while only 11.9% and 9.5% rated Excellent for the productivity agent and personalized narrative, respectively. Combining the top three
5.4 Results

Personalized Productivity Solutions

(a) Ratings of overall experience

(b) Average ratings across different criteria

Figure 5.6: Overall Experience: Figure (a) illustrates the participants’ overall experience ratings for three data presentation methods: Personalized Narrative, AI Productivity Agent, and Dashboard. Figure (b) compares the ratings of these methods across various criteria. We performed the Kruskal Wallis test to analyze the differences between the three groups in Figure (b), and criteria without any statistically significant differences (i.e., p-value > 0.10) has been marked with a crosshatch shading pattern - a series of intersecting lines drawn at an angle to create texture.

experiences (excellent, very good, and good), 81% considered the Dashboard good or above, 66.7% considered the AI productivity agent good or above, and 64.2% considered the personalized narrative to be good or above. Following the broader question, participants were asked to rate their experience on specific criteria for each method using a five-point Likert scale from Strongly Disagree (1) to Strongly Agree (5). In Figure 5.6b, a heatmap displays the average scores obtained by each method for these questions. Most darker blocks (in blue) are centered on the Dashboard, indicating higher average scores for the ease of exploring and understanding data, the use of the presentation method in real-life situations, and enjoyment in interaction. The boxes with crosshatch shading patterns (intersecting lines) indicate that the differences were not statistically significant (Kruskal Wallis test; p-value > 0.10) among the three methods for that particular criterion. The productivity agent scored better in only one area: engagement during data exploration. The personalized narrative
5.4 Results

Personalized Productivity Solutions performed the worst with the lowest scores among the three. Regarding anticipated usage frequency, we show in Figure 5.7a that 42% of participants selected weekly for dashboard and 30% selected weekly for productivity agent, which was the majority score for both groups. For personalized narrative, 28% participants said that they would use it weekly and the same amount also said they would rather use it monthly. 30% of the participants said they would use the personalized narrative occasionally. Lastly, participants ranked the three methods in order of preference (1 being the most preferred and their first choice, 3 being the least preferred and their last choice). As shown in Figure 5.7b, 55% rated the Dashboard as their most preferred method, followed by 27.5% preferring the productivity agent and 17.5% preferring the personalized narrative. Over half of the participants (52.5%) rated the personalized narrative as their least preferred method.

![Figure 5.7: Anticipated use and Ranking](image)

Figure 5.7: Anticipated use and Ranking: Figure (a) displays the anticipated usage of the three distinct data presentation methods. In contrast, Figure (b) showcases the participants’ rankings of these methods based on their preferences, with a ranking of 1 indicating the most preferred method and 3 signifying the least preferred method.
5.4.4. AI Productivity Agent Performance

In this section, we delve into the performance and capabilities of the AI productivity agent, examining the aspects of its interaction with users. We analyze the perceived capabilities of the agent, chat metrics, and productivity-related themes that emerged from users’ questions. Our goal is to gain insights into the effectiveness of the AI productivity agent.

Agent Capabilities. We designed the productivity agent based on the findings from our user survey and aimed to align it effectively with the insights from the first phase. To assess this alignment, we asked participants specific questions regarding the interactive chat agent’s perceived abilities. It is important to note that these questions were posed after participants had indicated their preferred mode of presentation, ensuring that their rankings were not influenced by these queries. For a comprehensive overview of the capabilities and participants’ ratings of the agent’s performance, please refer to Table 5.3.

29 participants (72.5%) agreed that the use of telemetry data in the agent enhanced its ability to provide personalized and empathetic support. To design the agent as knowledgeable, trustworthy, and transparent based on the user survey, 31 participants (77.5%) felt that the agent demonstrated these traits. Additionally, 34 participants (84%) agreed that the agent followed directions well. However, only 8 participants (20%) found the agent adaptive in its conversation style, with 17 (42.5%) remaining neutral and 11 (27.5%) considering it non-applicable. The agent performed well in avoiding negative traits identified in the user survey (we instructed the agent through its system prompt to avoid such negative traits), as 31 participants (77.5%) disagreed that it was overly critical, 33 (82.5%) disagreed that it was intrusive, and 33 (82.5%) disagreed that it was annoying, aggressive, or pushy. Nevertheless, the agent’s ability to see things from the user’s point of view received mixed responses,
Table 5.3: Perceived capabilities of the interactive agent. The responses are categorized as follows: SD for Strongly Disagree, SWD for Somewhat Disagree, NAND for Neither Agree Nor Disagree, SWA for Somewhat Agree, SA for Strongly Agree, and N/A for Non-applicable. We bold the group that has the majority of the responses – either Agree (SA, WA) or Disagree (SD, SWD) or Neutral (NAND, N/A).

<table>
<thead>
<tr>
<th>Capabilities</th>
<th>SD</th>
<th>SWD</th>
<th>NAND</th>
<th>SWA</th>
<th>SA</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>The agent provided useful tips and strategies for improving productivity</td>
<td>1 (2.5%)</td>
<td>1 (2.5%)</td>
<td>6 (15%)</td>
<td>15 (37.5%)</td>
<td>17 (42.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>The agent was able to understand my productivity challenges</td>
<td>0 (0%)</td>
<td>1 (2.5%)</td>
<td>8 (20%)</td>
<td>19 (47.5%)</td>
<td>9 (22.5%)</td>
<td>3 (7.5%)</td>
</tr>
<tr>
<td>The agent’s empathy and tone helped me feel more comfortable discussing my productivity concerns</td>
<td>2 (5%)</td>
<td>5 (12.5%)</td>
<td>10 (25%)</td>
<td>9 (22.5%)</td>
<td>12 (30%)</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>The agent personalized its suggestions based on my situation and needs</td>
<td>1 (2.5%)</td>
<td>2 (5%)</td>
<td>6 (15%)</td>
<td>16 (40%)</td>
<td>14 (35%)</td>
<td>1 (2.5%)</td>
</tr>
<tr>
<td>The agent provided practical solutions to overcome productivity obstacles</td>
<td>3 (7.5%)</td>
<td>1 (2.5%)</td>
<td>5 (12.5%)</td>
<td>16 (40%)</td>
<td>13 (32.5%)</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>The agent used a professional and direct language and tone</td>
<td>1 (2.5%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>13 (32.5%)</td>
<td>26 (65%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>The agent proactively suggested ways to improve my productivity</td>
<td>2 (5%)</td>
<td>0 (0%)</td>
<td>5 (12.5%)</td>
<td>17 (42.5%)</td>
<td>15 (37.5%)</td>
<td>1 (2.5%)</td>
</tr>
<tr>
<td>The agent was annoying, aggressive or pushy</td>
<td>26 (65%)</td>
<td>7 (17.5%)</td>
<td>3 (7.5%)</td>
<td>1 (2.5%)</td>
<td>2 (5%)</td>
<td>1 (2.5%)</td>
</tr>
<tr>
<td>The agent was being too intrusive</td>
<td>22 (55%)</td>
<td>11 (27.5%)</td>
<td>4 (10%)</td>
<td>2 (5%)</td>
<td>0 (0%)</td>
<td>1 (2.5%)</td>
</tr>
<tr>
<td>The agent was overly critical</td>
<td>20 (50%)</td>
<td>11 (27.5%)</td>
<td>5 (12.5%)</td>
<td>1 (2.5%)</td>
<td>0 (0%)</td>
<td>3 (7.5%)</td>
</tr>
<tr>
<td>The agent seems to be able to easily work out what I might want to talk about</td>
<td>1 (2.5%)</td>
<td>1 (2.5%)</td>
<td>10 (25%)</td>
<td>17 (42.5%)</td>
<td>10 (25%)</td>
<td>1 (2.5%)</td>
</tr>
<tr>
<td>The agent seems to have a difficult time seeing things from my point of view</td>
<td>8 (20%)</td>
<td>11 (27.5%)</td>
<td>12 (30%)</td>
<td>4 (10%)</td>
<td>3 (7.5%)</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>The agent seems to vary its conversational style to accommodate my mood or disposition</td>
<td>1 (2.5%)</td>
<td>3 (7.5%)</td>
<td>17 (42.5%)</td>
<td>7 (17.5%)</td>
<td>1 (2.5%)</td>
<td>11 (27.5%)</td>
</tr>
<tr>
<td>The agent seems to follow what I say and accurately reflects its understanding to me</td>
<td>1 (2.5%)</td>
<td>2 (5%)</td>
<td>3 (7.5%)</td>
<td>17 (42.5%)</td>
<td>17 (42.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>The agent demonstrated the traits of being knowledgeable, trustworthy, transparent, and responsive</td>
<td>1 (2.5%)</td>
<td>3 (7.5%)</td>
<td>5 (12.5%)</td>
<td>17 (42.5%)</td>
<td>14 (35%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>The agent’s use of telemetry data enhanced its ability to provide personalized and empathetic support</td>
<td>1 (2.5%)</td>
<td>5 (12.5%)</td>
<td>3 (7.5%)</td>
<td>17 (42.5%)</td>
<td>12 (30%)</td>
<td>2 (5%)</td>
</tr>
</tbody>
</table>

with only 19 participants (47.5%) disagreeing, 7 (17.5%) agreeing, and 12 (30%) remaining neutral. Other findings indicated that the agent proactively suggested productivity improvements (80%), provided practical solutions to overcome productivity barriers (72.5%), and offered useful tips and strategies (82.5%). Participants also felt that the agent personalized its suggestions based on their situation and needs (75%)
and understood their productivity challenges (70%).

**Productivity Conversations.** When examining the AI productivity agent interactions, we found that users exchanged an average of 10 messages. The average response time was 33 seconds. Users wrote an average of 9 words per message, while the chat agent provided more detailed responses with an average of 208 words per message. We examined the chat excerpts to better understand the aspects of the agent that resonated with users, as well as areas where improvements could be made to enhance the overall user experience. The majority of the messages were focused on productivity. In particular, we identified five main productivity-related themes from the users’ questions:

![Sample Conversation](image)

**Figure 5.8:** Sample Conversation: A sample response from the chat agent for the question, “What are my productivity barriers?”

(a) **Time Management and Work-Life Balance:** Users were concerned about managing their working hours, balancing work across different time zones, and un-
derstanding how their work habits compared to their peers. Questions in this theme included: “How is my work-life balance compared to peers?”, “How do you recommend balancing a team based in Pacific Time and Eastern Time with my Central hours?”, and “Do you notice a difference in work habits across the 4 weeks?”

(b) Meeting Efficiency and Management: Participants were interested in improving the efficiency and management of their meetings. They asked about meeting duration, attendance rates, and how to prevent meetings from running overtime. Example questions include: “Why are meetings shorter than 1 hour better for productivity?”, “How often do my meetings end on time?”, and “What recommendations do you have for ensuring meetings end on time?”

(c) Communication and Collaboration Habits: Users sought to enhance their communication and collaboration habits to boost productivity. They inquired about strategies to reduce email volume, refine collaboration habits, and improve communication with teammates. Sample questions are: “What can I improve in my communication habits to boost productivity?”, “How can I refine my collaboration habits for better teamwork?”, and “How can I be more interactive with my teammates?”. Figure 5.8 shows a sample response from the agent for one of these questions.

(d) Personalized Insights and Recommendations: Users aimed to receive tailored insights and recommendations based on their data to improve their productivity. They were interested in understanding their productivity barriers and receiving specific advice. Questions in this theme include: “What productivity barriers do I have?”, “What can I improve in my communication habits to boost productivity?”, and “What is one thing I should start doing to be more effective?”
(e) Tools, Templates, and Data Visualization: Participants expressed interest in tools, templates, and data visualization to streamline their work processes and better understand their productivity. They asked about email templates, note-taking templates, and ways to visualize their data. Example questions include: “Can you create email templates for common scenario email needs?”, “Please share a note-taking template and a template of a concise meeting summary with clear actions?”, and “Can you visualize this data for me?”. Note that the productivity agent we developed did not have the ability to create visualizations.

Section 5.5

Strengths, Weaknesses, and Improvement

Opportunities

To gain a deeper understanding of the pros and cons of each method and gather improvement suggestions, we asked participants to provide open-ended feedback. By considering participants’ insights and experiences, we can better understand the strengths and weaknesses of the AI productivity agent, visual dashboard, and AI-based personalized narrative, as well as explore opportunities for enhancing their overall performance and user experience.

Based on their feedback, we can say that participants appreciated the interactive chat agent for its ability to provide specific, tailored insights and recommendations. One participant mentioned, “The ChatBot was amazing!”, while another said, “I liked how it suggested solutions for my productivity without me even asking for it.” However, some participants found it labor-intensive and preferred using it as a secondary step after viewing a dashboard. A participant expressed this sentiment, stating, “I think the chatbot is a nice tool, but I would prefer to see the data in a visual form first, then maybe use the chatbot for follow-up questions.” Participants enjoyed the
dashboard presentation method for its visual appeal and ease of understanding. They expressed a desire for more interactivity and benchmarking functionality to compare their data against company or organizational averages. A participant commented, “I really loved the data visualization here, it’s exactly what was missing from the chatbot experience. I liked the colors too, they grabbed my attention.” However, another participant expressed a negative experience, saying, “While the dashboard looks nice, I feel overwhelmed with all the numbers and charts. I wish there was a simpler way to understand the data.” Similarly, another participant said, “The dashboard doesn’t provide adequate context for the data, leaving me unsure of how to apply the insights to my situation.” The personalized narrative method received mixed reviews from participants. Some found it succinct and helpful, while others thought it was too wordy and lacked visual appeal. One participant shared, “The text view is boring. I strongly prefer a more visual presentation method.” Another participant appreciated the actionable insights, stating, “It’s presenting tips without a lot of effort on my part.”

In summary, participants seem to prefer a combination of the dashboard visuals and the AI chat agent, offering both a visually appealing presentation and the ability to dive deeper into specific insights. Participants also expressed interest in using the data as a tool for having productive conversations with their managers and improving work-life balance. One participant suggested, “It would be super useful to me, especially as a manager to compare my results or my team’s results with others in my workgroup, org, company, etc.” To further improve the participant experience and the potential of the interactive chat agent, incorporating benchmarking, tracking changes over time, and enhancing the chat agent’s capabilities to display visuals could be beneficial.
Section 5.6

Discussion

Our study adopted a comprehensive two-phase approach to delve into user preferences for productivity agents and devise personalized solutions tailored to these preferences. The initial phase entailed a user survey to pinpoint the ideal persona for a productivity agent, examining individuals’ expectations and preferences regarding its features and behaviors. The subsequent phase centered on crafting an AI-based productivity agent, assessing its performance relative to alternative modes of telemetry data presentation, such as dashboard visualization and personalized narratives.

The user survey in phase one investigated various aspects of productivity, communication style, agent approach, personality traits, personalization, and privacy. The findings indicated that progress on scheduled tasks, resolving unexpected tasks, and assisting colleagues significantly contributed to participants’ sense of productivity. Most participants believed a productivity agent would be advantageous and that its personality would moderately or significantly influence its effectiveness. The participants desired AI agent assistance in areas like meeting summaries, documentation writing, start-of-day briefs, work summaries, and information retrieval. These results are in line with Kimani et al. (159), who discovered that emails and meetings were the primary distractions affecting information workers’ productivity. Both studies found that progressing or completing scheduled tasks was the top measure of productivity for participants. However, an intriguing difference emerged in the areas where respondents in the Kimani et al. study sought assistance. They primarily indicated a preference for help with reminders, task scheduling, breaking down larger tasks into manageable chunks, and avoiding distractions. While several of these aspects appear on our list of desired support areas, they are not ranked as highly as in the Kimani
et al. study. This discrepancy raises the question of whether the areas where people need assistance from AI productivity agents vary depending on their understanding of the agent’s capabilities. This notion brings up an interesting point that the responses in our study could have been shaped more by participants’ familiarity with the capabilities of LLMs, as portrayed by media and social discourse, rather than by their direct experiences with these technologies. There is a possibility that the hype and discussions about LLM-powered tools in public and online forums might have swayed participants’ perceptions, aligning them with tasks where AI is commonly viewed as beneficial. Our research did not explicitly separate preferences influenced by hands-on LLM use from those impacted by societal narratives. However, acknowledging this influence is critical, as it implies that user expectations for AI support can be shaped not only by personal use but also by the prevailing stories about AI’s potential in the wider world. This insight sheds more light on user attitudes towards AI productivity tools and their evolving expectations. It is important to note that the discrepancy in the areas where the respondents sought assistance could also stem from simply demographic variations in the research subjects. Kimani et al.’s study focused on information workers in positions such as program managers, administrative assistants, researchers, developers, engineers, and broadcast producers. Given that our participants report similar job roles, it’s unlikely that occupational differences are a major factor in the discrepancy but other demographic factors could have played a role.

In terms of communication style, participants preferred a communication tone that balanced casual and formal styles, with direct and concise messages. This preference might be attributed to the modern work environment’s emphasis on balancing professionalism and fostering a friendly atmosphere, which can promote collaboration and better working relationships. The preference for direct and to-the-point
messages might also be a result of the fast-paced nature of information work, where employees often have limited time to spend on interactions and prefer clear, concise information. The majority favored a proactive agent that displayed traits like being knowledgeable, trustworthy, and transparent, and took on the role of an assistant or expert rather than a supervisor. Regarding personalization and privacy, most participants preferred moderate to high personalization in the productivity agent and were comfortable sharing high-level telemetry data for better guidance. However, some participants expressed concerns about sharing personal information. These findings from the first phase provided valuable insights for designing a personalized productivity agent that effectively addresses users’ preferences and requirements.

Interestingly, approximately 23% of participants, as detailed in Table B.5 (Appendix B.8), expressed neutrality or did not believe productivity assistants would be beneficial for them. This subset highlighted several key concerns. One of the participants, for instance, differentiated between overall productivity and specific motivational challenges, noting that while they generally did not struggle with productivity, they occasionally faced difficulties in motivation, particularly with tasks that were unengaging or poorly defined. This gives us the impression that the participant believes such motivational challenges to be distinct from their overall productivity and not something they expect a productivity assistant to effectively address. Additionally, there were reservations about the factual accuracy of AI responses, especially for fact-based questions, reflecting a broader concern about the reliability of AI tools. Further, a notable apprehension was the potential over-reliance on AI and its impact on personal growth and skill development. Participants expressed uncertainty about their ability to challenge themselves if they became too dependent on AI assistance. Moreover, skepticism about the practical usefulness of AI in conversations was evident, with some participants emphasizing the need for real data to validate
the effectiveness of such tools. These insights suggest that while AI agents can offer structural support, they might be less equipped to address intrinsic motivational factors and that there is a need for continual improvement in AI reliability and user empowerment. Understanding these perspectives is essential for refining the design of AI productivity tools and enhancing their adoption and impact in the workplace.

In the second phase of the study, we developed the AI productivity agent based on the insights obtained from the initial user survey. We then compared the AI agent’s performance with two alternative methods of presenting telemetry data: a visual dashboard and an AI-based personalized narrative. User testing allowed for a comprehensive assessment of the AI productivity agent and the alternatives, considering factors such as task assistance, personalization, and overall user experience. The interaction data highlights the difference in user engagement and interaction patterns across the various presentation modes, indicating that users spent more time interacting with the AI productivity agent compared to the dashboard and personalized narrative. The AI productivity agent also offered more comprehensive responses, with a higher word count per message. The notable engagement duration with the AI agent, the longest among the tools evaluated, presents an interesting dimension to our findings. This prolonged interaction could be partly due to the AI agent’s capacity to provide more comprehensive and detailed responses, which inherently requires more time for users to read and digest. Such in-depth responses suggest a depth of interaction and a level of detail that may not be present in other tools, potentially leading to a richer user experience. However, it’s also important to recognize that some of this extended engagement might be attributed to the response generation time inherent in AI systems. Delays in producing responses, combined with the detailed nature of these responses, could have contributed to the longer interaction times. While longer engagement times do not unilaterally equate to higher productivity or satisfaction,
they do suggest a level of user interest and investment in the tool. This engagement could also reflect the users’ exploration of the AI agent’s capabilities, a positive sign of their willingness to integrate advanced technology into their productivity workflows. Future research should aim to untangle these factors – response time and response depth – to better understand their individual and combined impacts on user experience, engagement, and productivity.

The AI agent demonstrated promise in providing specific, tailored insights and recommendations. Most of the participants agreed that the use of telemetry data in the agent enhanced its ability to provide personalized and empathetic support, and they felt that the agent demonstrated desired traits such as knowledgeability, trustworthiness, and transparency. It is important to note that in addition to customizing responses, the availability of telemetry data can provide individuals with valuable insights into their work patterns and habits, fostering self-reflection. Previous research has demonstrated that the presence of such data can greatly enhance one’s ability to engage in self-reflection (209; 163). The agent also performed well in avoiding negative traits identified in the user survey, such as being overly critical, intrusive, and annoying. However, only a minority of participants found the agent to be adaptive in its conversation style. This could be because the study was not longitudinal, and participants only interacted with the agent for a few minutes. In such a limited time, they might not have had the ability to decide about the adaptiveness of the conversation style. Future studies could involve an extended interaction period, allowing users to better understand and evaluate the agent’s adaptiveness.

Participants also expressed a preference for a combination of the visual appeal and ease of understanding offered by the dashboard and the in-depth, personalized insights provided by the interactive chat agent. This preference suggests that users desire a comprehensive productivity tool that combines the strengths of both methods. One
possible reason for this preference is that visual representations of data can facilitate the easy comprehension and identification of trends, while the interactive chat agent can provide more detailed information and recommendations tailored to the user’s specific situation. There has been some prior research suggesting that visualizations can be better for browsing and getting oversights, while text summaries are better for targeted search for information (295). To further improve the interactive chat agent and better cater to user preferences, future work should focus on integrating the text and visualizations to address both needs, in addition to benchmarking, and other enhancements. Incorporating visual elements into the chat agent’s responses can help users more easily understand and interpret the data, catering to the preference for a combination of dashboard visualization and interactive chat agent. Benchmarking functionality, allowing users to compare their productivity data against their team, organization, or company averages, can provide valuable context and motivation for improvement. The personalized narrative method received mixed reviews, possibly due to the varying preferences of users when it comes to data presentation. Some may prefer a visually engaging presentation, while others might appreciate the succinct and informative nature of the narrative. This highlights the importance of offering multiple presentation methods or a hybrid approach to accommodate diverse user preferences. In summary, the study findings suggest that the interactive chat agent shows promise in providing tailored insights and recommendations, but there is room for improvement to enhance its effectiveness and user experience. For instance, in the example provided in Appendix B.9 Figure B.1b, there appear to be contradictions in the AI agent’s responses regarding email management strategies. Point 2 suggests specific slots for emails whereas Point 4 suggests spreading emails more evenly throughout the day. Such inconsistencies might have impacted participants’ perceptions and ratings of the agent, and highlight the need for further refinement in
the AI’s response generation process. Improving the prompts and the underlying logic of the AI could help ensure more coherent and consistent advice, potentially leading to better user satisfaction and trust in the AI agent. Future iterations of the tool will focus on enhancing the clarity and consistency of the AI’s recommendations. In addition, based on our findings, it seems that to create a more effective and user-centric productivity-enhancing tool, incorporating design principles like semantic zooming and progressive disclosure can be beneficial. Semantic zooming (291) allows users to view data at varying levels of detail based on their zoom level, presenting both visually appealing overviews and in-depth insights as needed. This could help users better understand relationships between productivity metrics and identify patterns or trends, promoting informed decision-making and productivity improvements. Progressive disclosure (239) presents information in a layered manner, initially offering essential information and revealing more details as needed. This approach caters to preferences for direct and concise messages, allowing users to control the level of detail they receive and ensuring the agent’s guidance remains relevant and useful. Integrating these design principles into the AI productivity agent may improve its effectiveness and user experience, striking a balance between visually engaging overviews and in-depth, personalized insights. Future research should consider the benefits of incorporating these principles into productivity agent development and explore additional techniques for creating effective and user-centric productivity-enhancing tools.

Our study also explored the capabilities of GPT-4 in enhancing the customization and delivery of productivity data from platforms like Viva Insights through an AI chatbot/agent. Its advanced natural language processing capabilities allowed us to transform standard productivity metrics into more personalized, interactive, and accessible advice. By leveraging the sophisticated context understanding of GPT-4, we were able to provide a conversational interface that offered more than just data visu-
alization; it provided tailored recommendations based on individual user patterns and preferences. This approach elevates the functionality of traditional productivity tools like Viva Insights, making the insights more actionable for the user. However, our study also recognized the importance of balancing personalization with user privacy. In designing our productivity agent, we adhered to privacy-by-design principles, using only high-level telemetry data and ensuring user consent for data usage. GPT-4’s efficiency in generating relevant and personalized responses without delving into sensitive personal details was key to maintaining this balance. This approach highlights the potential of LLMs to enhance user experience while respecting privacy concerns. Our study thus contributes to the growing body of research on AI-enhanced productivity tools, demonstrating the potential of advanced language models like GPT-4 in transforming how productivity insights are delivered and interacted with. Future research should aim to further isolate and examine the effects of presentation and functionality variations to provide clearer insights into the most effective ways of enhancing user productivity.

Despite the valuable insights gained from this study, some limitations should be considered when interpreting the findings and generalizing the results. One key limitation was the sample size and diversity, which may not have been large enough or sufficiently diverse in terms of demographics, professional backgrounds, and technology expertise to capture the full range of preferences, experiences, and productivity patterns among users. This lack of diversity could limit the generalizability of our findings to broader populations and contexts. The study’s duration may also not have been long enough to fully assess the effects of the personalized productivity agent, dashboard
visualization, and narrative summary, as participants may require more time to adapt to these tools and strategies. It is essential to acknowledge the potential influence of using different prompts on the generated outputs by GPT-4. As the AI model relies heavily on the input prompts provided, variations in these prompts can lead to dramatically different results, which could affect the overall consistency and reliability of the findings. The language model’s performance is highly sensitive to the phrasing and structure of the prompts, and even subtle changes might result in significantly different responses. Consequently, the outcomes derived from multiple iterations of prompts may not always be directly comparable, and this limitation should be considered when interpreting the results and drawing conclusions based on the AI-generated content. Also of note are the occasional hallucinations exhibited by GPT-4 and its poor ability to perform calculations accurately. As a result, we refrained from asking the agent to generate additional summary information or perform complex calculations, which could have potentially provided more insights into the productivity patterns. In our survey methodology (Appendix B.2), detailed questions about the interactive agent were followed by queries on the solutions’ strengths, weaknesses, and suggestions (Q28, Q29). This sequence could potentially induce a response bias, where participants’ views on the agent’s performance might have disproportionately influenced their assessment of its strengths and weaknesses. While we aimed for a comprehensive understanding of user experiences and randomized other aspects of the study including Q27 which explicitly asks for the benefits and drawbacks of each presentation method immediately after interacting with that method, future studies may benefit from randomizing the order of these questions to mitigate any potential bias.

The telemetry data extracted from Viva Insights provided valuable information on participant work patterns but did not capture a complete picture. Furthermore,
there might have been external factors or variables, such as changes in workload, organizational structure, or personal circumstances, that may affect productivity and thus are essential to be considered. In light of these limitations, it is essential to approach the study’s findings with caution and acknowledge that the results may not be universally applicable. Future research could address these limitations by employing larger and more diverse samples, extending the study duration, incorporating additional measures of productivity, and controlling for potential confounding factors. By doing so, researchers can continue to advance our understanding of the factors that contribute to productivity enhancement and the role of personalized productivity agents in this process. Finally, it’s important to consider the extent to which the AI productivity agent’s design was influenced by the user preferences identified in Phase 1. While there is a possibility that a closer alignment might have affected the comparative results, our design was a conscientious effort to balance user needs with technical feasibility. Future research could explore a more in-depth alignment of these insights, potentially offering a refined understanding of how well the AI agent meets user expectations. This ongoing process highlights the iterative nature of developing user-centric AI tools.

Section 5.8

Ethical Considerations

In this study, we adhered to ethical standards to protect the rights, privacy, and well-being of our participants. Informed consent was obtained from all participants, who were provided with a comprehensive overview of the study’s objectives, procedures, potential benefits, and risks. Anonymity and confidentiality were maintained by using unique identification codes for participant data, securely storing all information on encrypted servers, and limiting access to authorized research team members.
As the use of LLMs becomes increasingly prevalent in various aspects of our lives, it is essential to consider the ethical implications of integrating these models with other data sources, such as telemetry data. The combination of these technologies may lead to significant advancements in productivity and user experience. However, it also raises several ethical concerns, including data privacy, security, potential biases, surveillance, and the unintended consequences of automation. To address these concerns, we must maintain an ongoing conversation about the ethical implications of integrating large language models with other data. This includes addressing potential biases in AI models, ensuring transparency in data processing and decision-making, and establishing strong privacy and security protocols. It is also essential to involve a diverse range of stakeholders, such as ethicists, policymakers, and representatives from affected communities, in the development and deployment of these technologies. By fostering a proactive and responsible approach to the ethical challenges presented by the integration of large language models, we can harness the power of these technologies in a manner that respects individual rights and promotes the well-being of society as a whole.

Section 5.9

Conclusion

This paper presents a comprehensive study focused on understanding user preferences for productivity agents and designing personalized solutions to cater to these preferences. By conducting a user survey and analyzing the findings, we were able to identify key features and behaviors that contribute to user satisfaction and comfort when interacting with a productivity agent. Subsequently, we developed an AI-based productivity agent and compared its performance with alternative modes of telemetry data presentation, such as dashboard visualization and personalized nar-
The results from this study provide valuable insights into the factors that enhance productivity and user experience, as well as areas for improvement in the design and development of personalized productivity agents. By incorporating the lessons learned from this study, future work can focus on refining and optimizing productivity-enhancing tools and solutions, ultimately leading to improved efficiency and user-centric experiences for information workers.
Chapter 6

Contextual AI Journaling:
Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App

Section 6.1
Introduction

The significance of struggles with mental health among college students is becoming increasingly apparent, impacting students’ academic performance, social engagement, and overall personal development. Research, including findings from the American College Health Association (ACHA)–National College Health Assessment, highlights a concerning prevalence of anxiety, depression, and related issues among stu-
dents (18; 31; 221; 229). Students face a range of pressures, from academic challenges to social and personal hurdles, which affect not only their mental health but also their emotional resilience and personal growth (316; 304; 126; 334). While traditional mental health interventions administered by clinicians do provide personalized and context-specific support, emerging technologies present an opportunity to extend this support, making it more readily available, automated, and able to potentially overcome considerable institutional barriers. In addition, there is a need for innovative solutions that align with the digital habits of today’s students. We propose a novel study, MindScape, that integrates the traditional practice of journal writing with mobile technology and large language models (LLM) (227) to create a contextually-aware journaling application. The MindScape Android application benefits from on-device sensors and data to provide insights into the user’s daily life. It tracks aspects such as physical activity, social interactions, and location to form an understanding of the context in which the individual operates. By analyzing these data in real-time, the app can provide personalized, context-sensitive journaling prompts designed to provoke thought and reflection. The prompts aim to remind users to introspect and commit time to digitally record their thoughts, thus establishing regular self-reflection habits that are contextualized by their daily lives. MindScape represents a novel application class that incorporates behavioral intelligence into AI. We believe that integrating time-series data obtained from mobile phones and wearables, capturing real-time behaviors and patterns of users, with the capabilities of Large Language Models (LLMs) will give rise to a new category of AI applications driven by mobile sensing.

Journaling has long been recognized as a potent tool for self-reflection, enabling individuals to externalize thoughts, consolidate disjointed experiences, and identify patterns in their behavior and emotional states. This practice of regular introspection has been linked to a range of psychological benefits, from reducing distress symptoms
to enhancing overall wellbeing (297; 91). In this study, we explore the potential gains realized through the inclusion of personalization and context-awareness in journaling. The inclusion of personalization and context-awareness in journaling is more than just a technological novelty. It addresses certain inherent limitations in human introspection and memory recall abilities. People may not readily identify certain behavioral patterns or come to particular conclusions about their daily lives without some form of guidance or external input. This is where personalized and context-aware prompts can be valuable, as they may highlight aspects of users’ lives they may have overlooked. Additionally, human memory recall can be biased towards more recent experiences (recency bias) and peak emotional experiences (peak-recency bias), sometimes at the expense of equally significant past events (72; 165; 286). Context-aware journaling can help counteract this limitation by bringing forward relevant circumstances, events, or feelings from different timeframes in the users’ life. Lastly, by addressing these user limitations, personalized and context-aware journaling could not just improve the process of journaling, but also potentially enhance the mental health benefits associated with this practice.

Herein lies the novelty of our approach: using mobile sensing to capture behavioral data that reflects the user’s context and emotional state, and employing an LLM to generate journaling prompts that are highly relevant to the user’s current contextual situation and surroundings. To complement our context-aware journaling prompts, we introduce daily check-ins as a novel feature in our study app. These brief, simple texts are triggered four times daily and aim to encourage users to pause and reflect on their current experiences. For instance, a check-in might say, “Your morning seemed to include more than just tapping screens – a bit of chitchat too!” Users can respond with a quick thumbs up or thumbs down, allowing for a low-burden, high-engagement interaction. By leveraging contextual intelligence, our check-ins aim to increase user
attachment and engagement with the journaling app, while also making their reflections more meaningful and potentially amplifying the mental health benefits of journaling. The primary goal of these check-ins is to facilitate fleeting moments of self-reflection, helping users develop greater awareness of their thoughts, emotions, and behaviors throughout the day. By incorporating thumbs up/down responses, we simplify the reflection process, making it more accessible and increasing the likelihood of users engaging in regular self-reflection. Furthermore, the MindScape journaling app integrates additional contextual factors such as students’ mood while journaling, their academic stress levels, and temporal variables like weekdays or weekends. Early in our development, we conduct a qualitative user study with undergraduate students to understand their journaling habits and preferences. Insights from this study, revealing students’ desires for personalized, context-aware prompts aimed at fostering reflection on daily experiences, significantly influenced our app’s design. We believe our holistic approach allows for a more tailored and responsive tool, capable of providing meaningful support in the unique, often high-pressure, fast-paced environment of college life. Our paper makes the following contributions:

(a) We design MindScape – an AI-driven journaling app that integrates mobile sensing and LLM to deliver personalized, adaptive journaling prompts. We conduct an 8-week study with 20 college students to evaluate the efficacy of this system. By the end of the study, participants report up to an 11% improvement in well-being scores, with statistically significant enhancements in affect, loneliness, mindfulness, self-reflection, anxiety, and depression.

(b) We examine the check-ins and journaling prompts generated by the app, analyzing their similarity, topic coverage, and the frequencies of categories to which the prompts belong. We find that the morning check-ins often revolve around social and communication app usage, while afternoon check-ins shift towards
academic and social life experiences.

(c) We analyze linguistic differences between journals from contextual and generic prompts using the Linguistic Inquiry and Word Count (LIWC) (47). Our findings indicate that responses to contextual prompts exhibit more personal language, greater references to personal experiences and relationships, whereas broader emotional expressions (such as affect) are more prevalent in journals from generic prompts.

(d) We review participant feedback concerning their experience and the app’s usability and provide recommendations for future research. Overall, 85% of participants rate MindScape’s usability as good or excellent. Seventy percent consider the journal prompts to be moderately-to-very relevant, and 85% report that the contextual prompts sometimes, often, or always lead to more in-depth reflection compared to generic prompts, demonstrating the effectiveness of the MindScape app.

It is important to note that our objective is to introduce and evaluate a new journaling paradigm that integrates behavioral sensing and contextual awareness. This research conducts a proof-of-concept study on contextual journaling, specifically focusing on its effectiveness as a unique journaling method. We do not perform controlled trials to determine which approach is more beneficial. Ours is an exploratory study designed to potentially augment the classic benefits of journaling by utilizing the latest advancements in LLMs to provide an unobtrusive, effective tool for users to manage their wellbeing and growth. This approach is closely aligned with the Human-Computer Interaction (HCI) community’s interests, highlighting the significance of AI in enriching user-centric digital experiences. Bridging into Ubiquitous Computing (UbiComp), our research focuses on integrating these technologies into everyday routines. Our goal is for this tool to offer benefits and support and assist
students in developing lasting self-reflection and emotional mindfulness skills. We hope that this study will contribute significantly to the ongoing dialogue in HCI and UbiComp, particularly regarding the seamless integration of technology to enhance personal well-being, offering a comprehensive view of its practical application and user impact.

Section 6.2 Related Work

Journaling is a reflective practice where individuals record their thoughts, feelings, and experiences. The act of journaling promotes self-awareness (8; 341), processing of emotions (296), and cognitive organization of experiences (297). Studies have consistently shown that journaling can improve mood, provide stress relief, and overall, enhance mental well-being (297; 151; 218). As mobile devices and computers become more prevalent, they have reshaped the practice of journaling. The transition to digital journaling platforms brings conveniences that traditional paper-based methods lack. These include enhanced accessibility— ensuring that users can journal anytime and anywhere, heightened privacy—as entries are secured behind digital safeguards, and the ability to enrich journal entries with multimedia elements.

Journaling can be prompted or unprompted. Unprompted journaling allows for free expression without specific guidelines, giving users freedom to explore their thoughts and feelings. In contrast, prompted journaling uses specific questions or suggestions to guide the journaling process, providing a structure that can help focus and inspire the user. Such prompts are designed to encourage self-reflection, personal growth, and exploration of various topics and experiences. Several digital journaling platforms offer a wide range of prompts to initiate the writing and reflection journey, providing daily reminders to ensure users stay on track with their journaling. This
approach can be particularly helpful for users who are new to journaling or those looking to explore new areas of self-discovery and creativity. However, most prompted journaling applications rely on generic prompts not tailored to the user’s situation. Several studies demonstrate that question prompts are one of the main factors positively affecting reflection quality (62; 114; 60; 117). Thus, generic prompts, while useful, may reduce reflection quality due to their broad nature (16; 267).

Our study focuses on context-aware journaling, where journaling prompts are derived from behavioral data collected via smartphones. This approach enhances traditional journaling by offering prompts that closely align with users’ daily experiences and mental states. By using mobile sensing technology, capable of tracking activities, sociability, locations, and app usage, we generate dynamic prompts that reflect the nuanced aspects of an individual’s life. This approach differs from previous studies that have explored a broader range of personal informatics systems for reflection (68; 40), by integrating these insights into the journaling process. For example, Kocielnik et al. (163) leverage mobile based step count for reflection on activity level whereas Bakker and Rickard (21) use the MoodPrism app to help in mood tracking. Our method aims to mirror the reflective goals of various fitness and sleep apps and to offer deeper insights into users’ lifestyles and emotional patterns through personalized journaling. In addition, our study uses a wide range of contextual cues to facilitate journaling, a feature that sets it apart even from its closest counterparts like Apple’s journal application (14). While Apple’s offering leverages contextual data such as photos and location to generate prompts, our approach extends beyond conventional context-awareness to include an amplified set of signals such as: screen time; social, entertainment, and communication app usage; in-person conversations, calls, and text message exchanges.

Our study additionally integrates both sleep information, such as duration and
timing, as well as physical fitness metrics like activity levels, distance travelled, and time spent at the gym. Our study also considers location-based semantics like time spent in a cafeteria, Greek spaces, and other similar locations. This comprehensive approach sets our study apart by providing a more nuanced and detailed context for generating personalized journaling prompts. Our study also leverages LLM capabilities to enable the creation of intelligent, personalized journaling prompts. AI-driven tools have been used in therapy chatbots, virtual agents, and behavior change systems, offering personalized advice and support (67; 136; 347; 351; 182; 353; 174; 290; 238; 41; 154). These applications demonstrate the capacity of AI to understand and respond to a wide range of emotional and psychological states (208). Existing studies have leveraged LLMs for AI-mediated journaling (157; 158; 101). However, to our knowledge, none of the existing studies have integrated objective and passively observed behavioral data into AI-mediated journaling. By using an LLM framework to analyze behavioral data and generate relevant journaling prompts, we aim to investigate the potential for a nuanced, data-driven augmentation of the journaling process. Our study seeks to reinforce the benefits of journaling, while simultaneously exploring the effectiveness of context-aware prompts for highly reflective self-expression. Through this unique approach, we aim to optimize the impact of personalized digital journaling.

Section 6.3

Methodology

In this section, we detail our study methodology, which encompasses the study design, participant demographics, the mobile sensing behavioral data collected by our system, and the design of the personalized journaling prompts and check-ins.
6.3 Methodology Contextual AI Journaling

Figure 6.1: MindScape Study Design Overview: The study begins with a user research phase to capture insights into the journaling experiences and habits of students. The total duration of the study is 8 weeks, consisting of six weeks of contextual AI-based journaling prompts followed by two weeks of generic journaling prompts. We gather baseline and follow-up surveys at the start and end of the contextual journaling phase, respectively. Additionally, at the end of the 8th week, following the generic journaling phase, we collect a final set of surveys focusing on app experiences and overall study feedback. Note that the primary goal of our study is to explore the effectiveness of contextual AI-based journaling, not to compare its benefits with generic journaling. The generic journaling phase is included to provide a complementary analysis of linguistic differences in journals received for generic prompts versus contextual prompts, offering additional insights into the study’s outcomes.

6.3.1. Study Design

At the beginning of our study, we engage in a User Study focused on capturing participants’ perspectives using user-centered design principles. We conduct interviews with students to illuminate their needs and experiences with journaling, providing a foundational understanding for our research approach. The stages of the study are detailed in Figure 6.1. As we transition into the Development and Testing phase, we refine our methodology and initiate participant recruitment. We employ various channels such as posters, class-wide emails, Computer Science majors and minors email chains, student mailing lists, and collaborations with mental health-related campus clubs to reach potential participants. Out of 91 respondents expressing interest, 26
qualify for the study (the majority of them have Apple phones while our app only supports Android), with 20 ultimately signing the consent form to participate. We take care to exclude individuals with high depression scores, as indicated by elevated Patient Health Questionnaire-8 (PHQ-8) survey results, to ensure safety due to the unmoderated nature of the reflection prompts. Once enrolled, participants install the MindScape Android app on their phones. The central six weeks of our study involve participants interacting with contextual AI-driven journal prompts delivered through the app. This begins with an onboarding process, where participants complete a baseline survey that captures their initial journaling habits, demographic details, and psychological states via standard surveys focused on well-being and self-reflection. At the conclusion of this six-week contextual journaling phase, we administer a follow-up survey using the same standard questionnaires. This enables us to gauge changes in well-being, personal growth, and reflection, assessing whether AI-driven contextual journaling contributes positively to participants’ development. Additionally, we conduct weekly Ecological Momentary Assessment (EMA) – a research methodology that involves repeatedly collecting self-reported data form participants to capture dynamic changes and patterns over time – to monitor changes in participants’ well-being and reflection over-time. Please see Appendix C.1 for the list of surveys and questions we ask participants.

Following the initial contextual journaling phase, the participants enter a two-week period of generic journaling, receiving a uniform prompt via the MindScape app: “What’s on your mind today? Use this journal entry to explore freely any thoughts, feelings, memories, or experiences—anything you’d like.” Due to a limited sample size, a full randomized controlled trial was not feasible. Nevertheless, this phase provides an opportunity to compare and contrast traditional journaling with our AI-driven contextual method. After completing the full eight-week study
duration, participants receive the final study feedback survey, which collects their insights on their journaling experience. This includes thoughts on the app’s usability and performance as well as any additional feedback or suggestions. Participants are compensated up to USD 130 for their involvement. The study has received approval from Dartmouth College’s Internal Review Board, ensuring all procedures meet ethical standards.

6.3.2. Demographics

We recruited 20 students from Dartmouth College for our study. Out of these participants, a majority, 60% (N=12), identify as female, while 35% (N=7) identify as male, and one participant (5%) identifies as non-binary. The cohort comprises 12 graduate students and 8 undergraduate students. When examining racial demographics, 35% (N=7) of participants identify as White or Caucasian, 25% (N=5) as Asian, 20% (N=4) as Black or African American, 15% (N=3) report belonging to multiple racial categories, and 5% (N=1) report ‘Other’. Age distribution among the participants shows that 65% (N=13) are within the 18-24 age bracket, 30% (N=6) fall into the 25-34 age range, and 5% (N=1) are 45 years old or above. Regarding journaling experience, 55% (N=11) of the participants currently maintain a journal, 20% (N=4) do not keep a journal at present though they have journaled in the past, and 25% (N=5) have never engaged in journaling.

Table 6.1: Behavioral Data Categories: Users are required to prioritize among the four behavioral data categories, each encompassing specific feature sets.
6.3.3. Mobile Sensing based Behavioral Data

The MindScape app automatically infers user activities, like movement and rest, analyzes conversation lengths, and gathers data on screen usage and location (see Table 6.1). This provides an integrated view of a user’s daily patterns, social interactions, and digital habits. For example, the sensing data might reveal patterns in how often participants attend social functions, dine at campus facilities, or go to the gym. This information allows us to tailor the journaling prompts to align with the participant’s current experiences and to support their emotional well-being. As part of gathering this data, we create a semantic map of the college campus, with locations such as dining areas and gyms marked, allowing the app to accurately infer the context of participants’ activities. This allows for prompts to be customized, encouraging reflection on particular events of the day. The integration of the GPT-4 LLM enables the translation of this rich, multi-faceted behavioral data into personalized and contextually relevant journaling prompts and frequent check-ins that enhance positive introspection and participant engagement. All data collected are temporarily stored on the participant’s phone and then securely uploaded to the MindScape cloud. We then leverage the GPT-4 model through OpenAI’s API (247), allowing us to process the collected behavioral data and additional contexts to generate tailored prompts. Addressing potential concerns relating to participant privacy, we ensure all data sent for processing via OpenAI’s GPT-4 model are de-identified and consist only of high-level metadata. This approach includes stripping any potentially personally identifiable information before the data is utilized to generate tailored prompts. We acknowledge that a locally hosted open-source model could offer an alternative to mitigate privacy concerns further, albeit with possible performance trade-offs. In this study, our focus is oriented towards understanding the potential and efficacy of this novel application of AI in journaling practices. Given this emphasis, we decided to
utilize OpenAI’s GPT-4 model for its robust performance and scalability capabilities.

Figure 6.2: MindScape App Workflow: Users sign in, set preferences in four categories, and start journaling with a notification tap. Journaling workflow includes an emoji based mood assessment, a one-minute breathing exercise, followed by contextual prompts. The final screen showcases the daily context-aware check-in.

6.3.4. Personalized Journaling Prompts

Upon installing MindScape app, participants were prompted to allow the app permission for data collection. Then, they rank their journaling interests in four key areas — Social Interaction, Sleep, Digital Habits, and Physical Fitness. We identify these four key areas through interviews with students on campus (See Section 6.4.1).
Because we collect many different types of data, we want to ensure the journaling prompts we provide are actually helpful to participants. Thus, we use these categories to identify what matters most to each individual participant. We also include the user’s preferences (i.e., category ranking) in the prompt for GPT-4 (246) to generate more relevant journaling prompts. During their enrollment, each user provides us with their usual bedtime for both weekdays and weekends. Journaling notifications are triggered two hours before their reported bedtime. When a notification is tapped, participants are redirected to the app’s journaling screen. There, they are first asked how their day was, followed with a one-minute breathing exercise, and finally, they are asked to write or record (i.e., audio) their journal entry. Only at this point can the participants see the personalized journaling prompt. Participants can also open the app and journal whenever they prefer. Note, the one-minute deep breathing exercise before journaling is based on findings that short relaxation techniques can improve mental clarity and emotional readiness (36; 356). This step aims to help users transition to a reflective mood, enhancing their focus for more insightful journaling. It is intended to make the journaling process a calming, enriching routine. Figure 6.2 shows different screens of the application.

Contexts The GPT-4 prompt composition process incorporates several layers of contextual data:

- **Personal Priorities:** The user’s preferences across the four journaling categories ensure that the journaling prompts mirror individual interests.

- **Prompt Variability:** The system ensures that new prompts are different from the previous two, generating diverse and engaging content.

- **Temporal Data Analysis:** Behavioral data from weekdays are contrasted with a 30-day historical average to establish context. On Saturdays, the app
encourages users to reflect on general themes from the preceding week, rather than daily behaviors (for example, “Recall a recent academic success. How did you achieve it and what did it teach you about your resilience or strategy?”). Sundays are used for a comprehensive review including additional data points—such as Greek house attendance and sleep quality—to capture weekend patterns pertinent to college life. \textit{Note: In the U.S., ‘Greek houses’ are fraternity or sorority residences, where social and organizational activities are hosted.}

- \textbf{Academic Calendar Awareness:} As the academic term structure influences stress, the current week of the term is considered during prompt generation, intending to offer supportive content during high-stress phases.

- \textbf{Mood Consideration:} If a participant reports a low mood, GPT-4 is prompted to offer journaling prompts that evoke self-compassion or gratitude—strategically fostering a nurturing journaling environment. By guiding users towards reflecting on aspects they are grateful for or encouraging kindness towards themselves, the hope is that these prompts can shift focus from negative thoughts to more positive, affirming ones. It is a strategic, evidence-based approach aimed at offering immediate emotional relief while contributing to long-term emotional well-being, resilience, and mental health (87).

Our methodology emphasizes customization, employing both user preferences and behavioral signals to empower participants in their reflective journaling practice. In Figure 6.3, we show how all these come together to form the input to the GPT-4 LLM.

6.3.5. Context-aware Check-ins

The check-ins are “micro context-aware nudges” based on users’ data, and are answered with a quick thumbs up or thumbs down response. For example, “Caught up
6.3 Methodology Contextual AI Journaling

Figure 6.3: Prompt Template for Weekday Journaling: The input prompt to GPT-4 is composed of four parts: 1) System prompt 2) User context 3) Rules to optimize the prompt and 4) The strategy to generate the journaling prompt.

with some calls and social apps this morning - digital world kept you busy, I bet!“. The MindScape app offers such check-ins four times a day at 12.30 PM, 3.30 PM, 6.30 PM and 11 PM. These times are strategically selected to suit the daily rhythms of college students, ensuring the interaction remains brief and unobtrusive.

Each check-in is designed to incorporate the behavioral data gathered during the time period extending from the previous check-in up to the current one. For instance, the 3:30 PM check-in uses data collected from 12:00 PM to 3:30 PM, while the 6:30 PM check-in uses data gathered from 3:30 PM to 6:30 PM. This approach ensures that each check-in is responsive to the most recent behavioral data captured for the participant. The goal of these check-ins is to both increase the visibility of the app (as opposed to users seeing it just once a day for journaling) as well as to increase reflection on behavior through a casual, quick touchpoints. Please refer to Appendix C.2 for the complete GPT-4 prompt we use to generate check-ins. Important to note: the
responses to check-ins (i.e., thumbs up/down) are not utilized as feedback to inform the GPT-4 model or processed further to influence subsequent prompts. They serve solely as a simple engagement mechanism for users.

Section 6.4

Results

In the following section, we present the results from our study. We begin by examining the journaling prompts and check-in messages, followed by an analysis of the linguistic content of the journals, including a comparison between contextual and generic journals. We then review the changes in well-being and personal growth scores, as observed in the follow-up survey conducted after the study. Finally, we discuss participant usability and feedback, and offer recommendations for future researchers.

6.4.1. User Study

We conduct qualitative user studies through in-depth interviews with five undergraduate students at Dartmouth College, with the goal of understanding their journaling habits, preferences, and expectations for potential personalized prompts that could be generated by the MindScape app. The participants, aged between 18-24 and comprising 3 males and 2 females, are selected through targeted invitations extended by our team to ensure a range of insights into the efficacy and impact of personalized journaling within the university context.

During these interviews, students are introduced to the various types of data that could be captured via their smartphones. Based on the signals that we can feasibly track, such as location data, physical activity, app usage, and others, students identify four main areas of interest that they believe would be most beneficial for personalized journaling prompts. These preferences include:
• **Social Interactions:** Reflecting on social activities and relationships, influenced by data on in-person conversations, phone logs and time spent at different locations (such as fraternities, social places)

• **Sleep Patterns:** Insights derived from sleep tracking data to encourage better sleep habits and reflections on the impact of sleep on daily functioning.

• **Physical Fitness:** Using activity tracking data, and time spent at the gym to monitor progress, and reflect on the connection between physical health and overall well-being.

• **Digital Habits:** Observations on app usage and screen time to encourage healthier digital interactions and balance.

Participants also share their preferable contexts and times for engaging with the app—highlighting a tendency to journal during quieter moments of the day, or when experiencing stress, suggesting that prompts should be adaptive to their emotional states and academic schedules.

• **Motivations and Barriers:** Participants note journaling as a helpful tool for emotional processing and stress management. However, common barriers cited include uncertainties about what to write, time constraints, and inconsistent journaling habits. These insights underline the opportunity for MindScape to incorporate features like structured prompts and integrated reminders to help users navigate these challenges.

• **Adaptive Features:** There is a strong interest in receiving journaling prompts that adapt based on their sensed emotional state or specific stressors, such as exam periods or significant personal events, demonstrating the need for adaptive AI functionalities within the app.
• **Self-reports and Sensing-based Prompts:** Participants are willing to provide self-reports at different times of day, suggesting that multiple daily check-ins and end-of-day journaling are feasible. They respond positively to location-specific prompts, such as those related to meals in the cafeteria or academic work in the library.

• **Academic Stress:** The user study validates our understanding of academic stress among students, particularly highlighted during periods like exams and project deadlines. Students report increased stress levels impacting their sleep, social interactions, and overall well-being.

These insights are instrumental in tailoring the development of the MindScape app’s prompting mechanisms. We integrate additional contextual factors such as awareness of the academic calendar, mood fluctuations, and personal priority tracking to enrich the user’s engagement with reflective practices meaningfully. With these adaptive and user-centric features, the MindScape app aims to enhance users’ well-being through tailored, data-informed interactions.

### 6.4.2. Contextual Journaling Prompts and Check-ins

After conducting the MindScape study, we have a total of 661 journaling entries across 8 weeks of the study - 533 in the first six weeks of contextual journal prompts and 128 in the last two weeks of generic journal prompts. During the initial six weeks of contextual journal prompts, participants engage for an average of five weeks, submitting an average total of 26.65 entries per participant. In the subsequent two weeks of generic journaling, participants submit journals for an average of 1.72 weeks with 7.11 entries. Overall, participation across the entire study spans an average of 6.5 weeks with 33.05 entries per participant. Similarly, for check-ins, participants respond to a total of 2,985 check-ins throughout the study, with an average per participant of 36.85
in the morning, 44.15 in the afternoon, 43.15 in the evening, and 26.55 at night. We send participants four check-in messages each day, and responses are predominantly received for the afternoon and evening check-ins. We track participants’ engagement through ‘thumbs up’ or ‘thumbs down’ responses at each daily check-in. The night check-ins have the least participation likely due to their timing at 11 PM but show the most favorable response with a ratio of 4.8 thumbs up for every thumbs down (436 to 90). Morning check-ins also demonstrate strong positive feedback with a ratio of 4.2 (590 to 140). The afternoon sessions indicate robust engagement with a ratio of 3.4 (676 to 199), while the evening sessions have the lowest positive ratio among the sessions, standing at 3.1 (647 to 207).

At the beginning of the study, we offer participants the opportunity to personalize their experience through the MindScape app by prioritizing journaling categories based on their individual goals. The categories are: Social Interaction, Sleep, Physical Fitness, and Digital Habits. Upon installing the app, participants rank these categories in order of importance, tailoring the contextual journaling prompts they receive during the first six weeks of the study. The order of these categories are randomized when displayed to the user, and participants re-order the categories to rank them according to their preferences. Figure 6.4a visually presents the distribution of participant preferences across four priority ranks. A clear preference for Social Interaction emerges, with seven participants ranking it as their top priority and eight as their second. Digital Habits follow as the second most prioritized category, with six participants identifying it as their primary focus and four as their secondary. Sleep ranks third, with five participants prioritizing it first and three second. Physical Fitness is the least prioritized category, with only two participants identifying it as their primary focus and five as their secondary, and seven ranking it last. Note that some prompts do not fit neatly into one category, as they combine elements of multiple
domains (e.g., “Reflect on how your current digital habits are contributing to better sleep and how you might further enhance this balance.”). Please see Appendix C.4 for examples of the contextual journaling prompts delivered by MindScape.

Figure 6.4: Prompt Categories and their Distribution: Figure (a) illustrates the participants’ preferences among four categories, with Social Interaction emerging as the top priority for the majority. Figure (b) displays the distribution of journaling prompts received by the participants. This distribution aligns with the ranking of categories chosen by participants in Figure (a), with the exception of Sleep, which occurs least frequently.

Figure 6.4b displays a pie chart showing how the MindScape app distributes journaling prompts across four priority categories, aligning with participants’ prioritization. A significant 42% of generated prompts focus on Social Interactions, reflecting its top priority among 15 participants (75%). Digital Habits, ranked next most important by participants, account for 22% of prompts. However, the remaining prompts reveal an unexpected shift: 15% focus on Physical Fitness, surpassing Sleep, which comprises only 7% of prompts. This is likely because Physical Fitness encompasses a broader range of signals, including daily exercise, distance traveled, and physical activities like standing, running, and walking. In contrast, Sleep considers only two signals: total sleep duration and schedule. With a wider array of signals, there is greater potential to highlight changes or improvements in more signals compared to just two for sleep. Additionally, sleep patterns tend to be more stable throughout the
study period for participants, which means they would not be frequently highlighted by the GPT-4 prompts we designed. Recall that in the prompts to GPT-4, we indicate whether a particular signal was increasing, decreasing, or remaining stable, and by how much. Furthermore, 14% of prompts do not align directly with any of the four primary categories; these broader prompts, introduced on weekends, diversify the journaling experience and encourage reflection on a wider range of topics beyond core areas identified by participants.

Figure 6.5: Check-in Prompt Topics: We employ topic modeling to analyze the various concepts addressed in the check-in prompts. These prompts cover a diverse range of topics that vary depending on the time of day. For example, the daily usage of social and communication apps is prominently featured in the morning but least common in the night.

We understand the categories of our journaling prompts because we ask participants to rank their priorities, which we then pass to GPT-4. However, to better understand the content of our check-ins we perform topic modeling on the check-in
prompts. The process begins by extracting embeddings using the `all-mpnet-base-v2` sentence-transformer model (139), a robust model trained on large sentence-level datasets and fine-tuned with 1 billion sentence pairs. We create numerical representations of the documents through these embeddings and proceed to reduce their dimensionality using Uniform Manifold Approximation and Projection (UMAP) (210). UMAP effectively preserves both local and global structures of the dataset, maintaining crucial information that aids in clustering semantically similar documents. Following dimensionality reduction, we cluster our data using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (55). This technique not only identifies clusters of various shapes but also excels in pinpointing outliers, thus ensuring that documents are not inappropriately forced into clusters, enhancing the accuracy of the topic representations by minimizing noise. Next, we tokenize the topics and employ the class-based term frequency-inverse document frequency (c-TF-IDF) algorithm to create detailed topic representations. The c-TF-IDF method efficiently generates precise topic descriptions by selecting sets of words that best typify a collection of documents. After generating these initial topic representations, we refine them to produce final labels. We select candidate keywords and documents that most effectively describe each topic, according to the c-TF-IDF. These are then submitted to GPT-4, which is tasked with generating output that closely aligns with the specified topics. Throughout this procedure, we leverage the BERTopic (124) python library, an advanced tool designed to streamline and optimize topic modeling tasks.

Figure 6.5 displays the top 10 topics identified through topic modeling, sorted according to the time of day when the check-in prompts were issued: morning, afternoon, evening, and night. This organization offers insights into the contextual relevance of each topic to the students’ daily routines, as reflected in their interac-
6.4 Results

Contextual AI Journaling

tions with the prompts. In the morning, there is a predominance of prompts related to *Daily Usage of Social & Communication Apps*. It’s likely that there are few other significant activities in the morning hours (6-11:45 AM) other than students engaging in communication via their phones or still being in their dorms, presumably sleeping, so the topics predominantly revolve around these aspects. As the day progresses, however, the nature of the highlighted topics shifts, reflecting changes in students’ focus and activities. For instance, topics identified during the afternoon, such as *Study Spot Experiences*, increase significantly, likely indicative of students attending classes, working, or visiting libraries. *Dorm Life & Social Interactions* related prompts also see a rise during the afternoon and peak in the evening, likely due to increased proximity to peers and social interactions. Interestingly, *Daily Usage of Social & Communication Apps* reaches its lowest point during the evening prompts, whereas *Engaging Conversations Throughout the Day* hits its peak. At night, *Daily Usage of Social & Communication Apps* increases again from its evening lows, but *Study Spot Experiences* decreases, likely because students are winding down. Note that since we have a lower response rate for the nighttime prompts, it possibly influences the limited topics we identify for nighttime prompts. Please see Appendix C.5 for a sample of check-ins delivered by the MindScape app.

We generate both check-in and journaling prompts using multiple signals. Although we instruct the LLM to avoid generating similar prompts, we now try to assess the effectiveness of this strategy. We analyze the repetitiveness of the prompts by leveraging the embeddings we previously created using the *all-mpnet-base-v2* sentence-transformer model. We then employ cosine similarity in two distinct analyses to compare the semantic similarity of these embeddings. By calculating the cosine of the angle between two non-zero vectors in a multi-dimensional space, cosine similarity effectively quantifies the similarity between document vectors. Scores range from -1
6.4 Results

We compare the similarity of each prompt with every other prompt to evaluate the overall diversity in the language of the generated prompts. Additionally, we assess short-term repetitiveness by comparing each pair of consecutive prompts, ensuring the prompts are not annoying to users and reflect day-to-day variability in the data. This analysis is crucial because frequent replication of prompts, such as identical journaling prompts over two days or repetitive afternoon and evening check-in messages, could negatively impact the app’s performance and user retention. In Figure 6.6, we present the similarity scores for the journaling prompts and check-in messages.

Figure 6.6a displays a boxplot of the overall similarity scores (Y-axis) for contextual journaling prompts, showing an average score of 0.38 (indicated by the brown circle and dotted lines). The average minimum similarity among journaling prompts is 0.33, while the average maximum is 0.46. An overlaid violin plot reveals that most average similarity scores cluster around 0.36, with the width of the violin curve corresponding to the approximate frequency of data points in each region, peaking at 0.36. Similarly, Figure 6.6b displays a box plot for the range of similarity scores for each user, showing that nearly all participants have similarity scores below 0.5 in most cases, with averages consistently below 0.4. One exception is Participant 13, who experienced a technical issue causing delayed problem identification and received the same prompt for several days, likely resulting in their elevated similarity score.

Similar to the journaling prompts, Figure 6.6c and 6.6d display the within-check-ins similarity and consecutive check-ins similarity, respectively. We assess the semantic similarity of check-ins within each period—morning, afternoon, evening, and night. In Figure 6.6c, we observe that morning check-ins tend to have a slightly higher similarity score compared to other periods, likely due to fewer variable signals during
6.4 Results

Contextual AI Journaling

(a) Journaling Prompt Similarity (Comprehensive Comparison)

(b) Journaling Prompt Similarity (Sequential Comparison)

(c) Check-in Similarity (Comprehensive Comparison)

(d) Check-in Similarity (Sequential Comparison)

Figure 6.6: Repetitions of Prompts and Check-ins: We analyze the semantic similarity among journaling prompts and check-ins. Figure (a) displays the similarity across all journaling prompts, while Figure (c) shows the similarity among all check-in prompts within specific time periods. Conversely, Figure (b) and Figure (d) illustrate the similarity between each consecutive pair of journaling and check-in prompts, respectively. The analysis reveals that check-ins tend to be more repetitive than journaling prompts.

these hours, such as app usage or the participant being in the dorm and presumably sleeping. Night check-ins show a wider range of similarity scores, from 0.34 to nearly 0.6, indicating varied experiences in terms of similarity among different participants during this period. Conversely, evening check-ins appear to have a narrower spread, suggesting more consistency in the similarity of check-ins for all participants during these hours. On average, we find that check-in prompts are more similar (above
0.46) than journaling prompts (0.38), which could be attributed to the limited size of the signal sets. This limitation may not significantly impact journaling, which occurs once a day, but becomes more problematic for check-ins that occur four times a day. In Figure 6.6d, we see a per-user breakdown of the similarity scores of consecutive pairs of check-ins received by the participants, with an average similarity score above 0.4, compared to under 0.4 for journaling prompts. Several instances show tails reaching 0.6 and higher, indicating more frequent repetition in check-ins in terms of semantic similarity compared to journaling prompts. While a score of 0.38 for journaling prompts indicates some degree of similarity, they are not highly similar. However, given the higher scores observed in check-in prompts, we identify room for improvement to reduce repetitiveness and enhance the novelty and relevance of each check-in.

6.4.3. Journaling Responses Deeper Dive

In this section, we dive deeper into the journal entries submitted by participants. We analyze and compare responses to both contextual and generic prompts. Additionally, we assess the contextual prompts specifically designed to foster self-compassion and gratitude when participants report low moods.

Journaling Showdown: Contextual vs. Generic: For the initial six weeks of the study, the MindScape app sent contextual journaling prompts that were dynamic and tailored day-to-day based on the participants’ passively collected behavior. After this period, the subsequent two weeks featured generic, static prompts that consistently asked, “What’s on your mind today? Use this journal entry to explore freely any thoughts, feelings, memories, or experiences – anything you’d like.” We now compare participants’ responses to these generic prompts with those to the contextual prompts.

For clarity, we refer to entries from contextual and generic prompts as “contextual journals” and “generic journals”. We list the most frequent unigrams and bigrams in
Table 6.2: Top Unigrams and Bigrams: The table below presents the most frequently occurring unigrams and bigrams found in contextual and generic journal entries.

<table>
<thead>
<tr>
<th>Unigram</th>
<th>Frequency</th>
<th>Bigram</th>
<th>Frequency</th>
<th>Unigram</th>
<th>Frequency</th>
<th>Bigram</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>204 (26.47%)</td>
<td>little bit</td>
<td>27 (4.13%)</td>
<td>today</td>
<td>77 (50.00%)</td>
<td>good day</td>
<td>11 (8.59%)</td>
</tr>
<tr>
<td>work</td>
<td>131 (18.53%)</td>
<td>feel like</td>
<td>23 (4.13%)</td>
<td>things</td>
<td>33 (17.19%)</td>
<td>little bit</td>
<td>7 (5.47%)</td>
</tr>
<tr>
<td>today</td>
<td>126 (17.20%)</td>
<td>screen time</td>
<td>22 (3.38%)</td>
<td>work</td>
<td>29 (17.19%)</td>
<td>feel like</td>
<td>7 (4.69%)</td>
</tr>
<tr>
<td>people</td>
<td>114 (17.39%)</td>
<td>spend time</td>
<td>16 (3.00%)</td>
<td>week</td>
<td>23 (9.38%)</td>
<td>today got</td>
<td>6 (4.69%)</td>
</tr>
<tr>
<td>phone</td>
<td>92 (12.10%)</td>
<td>new places</td>
<td>13 (2.06%)</td>
<td>tomorrow</td>
<td>19 (12.50%)</td>
<td>looking forward</td>
<td>6 (4.69%)</td>
</tr>
<tr>
<td>week</td>
<td>92 (12.10%)</td>
<td>makes feel</td>
<td>12 (2.06%)</td>
<td>feeling</td>
<td>18 (13.28%)</td>
<td>even though</td>
<td>5 (3.91%)</td>
</tr>
<tr>
<td>friends</td>
<td>80 (13.80%)</td>
<td>every day</td>
<td>12 (1.69%)</td>
<td>nice</td>
<td>16 (11.72%)</td>
<td>next week</td>
<td>5 (2.34%)</td>
</tr>
<tr>
<td>things</td>
<td>58 (9.45%)</td>
<td>social media</td>
<td>11 (1.88%)</td>
<td>bit</td>
<td>15 (10.94%)</td>
<td>felt like</td>
<td>5 (2.34%)</td>
</tr>
<tr>
<td>nice</td>
<td>52 (8.13%)</td>
<td>less time</td>
<td>11 (1.88%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

both journals in Table 6.2, where we see some distinctive and some similar patterns in the language used across these two types of journal entries. In contextual journals, “time” emerged as the most frequent unigram, appearing in 26.47% of entries. This was closely followed by “work” and “today”, indicating a focus on daily activities and schedules. The higher occurrences of bigrams such as “little bit” and “feel like” suggests reflections on experiences and emotions. The usage of terms related to social activities and digital habits, such as “friends”, “social media”, “spend time”, “phone”, and “screen time”, in contextual journals hints at the specific nature of these prompts i.e., concern for digital habits and social interaction. Conversely, in generic journals, “today” was the top unigram, found in 50% of entries, suggesting a reflective focus on the day’s events. This is followed by “time” and “things”, highlighting discussions around daily occurrences and tasks. Both types of journals share common terms such as “time,” “today,” and “work,” meaning possibly that there is shared themes across all journal entries. This might suggest that despite the differences in prompts, participants might reflect on daily experiences, tasks, and relationships.

While we are familiar with the primary categories and associated themes of the contextual journals, the generic journal prompts invite participants to write freely about anything they wish to explore. Hence, to deepen our understanding of the thematic content within these generic journals, we employ the same topic modeling
6.4 Results

The approach described in the previous section led to the identification of four primary topics characterizing participants’ generic journals: Daily Experiences, Daily Activities, Productivity Management, and Academic & Personal Growth. We present this in Figure 6.7. As we can see in the figure, the topic of daily experiences dominates 40 journals, rich with words like “happy,” “today,” “stressed,” and “progress,” revealing a wide emotional range. A representative journal illustrates this diversity: a participant recounts feeling conflicted, missing home, and feeling slightly...
unwell, juxtaposed with the joy of new friendships and a hopeful outlook toward future opportunities for openness and connection. Encounters with daily activities are described in 38 entries, encompassing descriptions of routine or special events, highlighted by words such as “exciting,” “fun,” “enjoy,” and “conversations.” One journal entry particularly captures the essence of an extraordinary day spent skiing with friends, encapsulating the exhilaration and contentment derived from engaging in an activity in good company. Productivity management is the focus of 35 journals, which centers on the pursuit of efficiency and accomplishment, underscored by terms like “productive,” “working,” “progress,” and “rest.” A specific journal describes a day spent working from home, balancing productivity with self-care, and reflecting on the importance of not feeling guilty for taking breaks. The least represented, with 15 journals, is the topic of academic & personal growth, revolving around advancement in academic endeavors and personal development, with words like “progressing,” “studying,” “internship,” and “fellowship.” One standout journal entry describes attending an inspiring symposium, which increased the participant’s confidence in their research and its potential impact.

Following this, we utilize the Linguistic Inquiry and Word Count (LIWC) to further compare the journal entries. LIWC is a research tool that provides insights into the psychological and emotional underpinnings of language use. By analyzing the frequency of psychologically meaningful words, LIWC allows us to understand aspects such as emotionality, social relationships, and thinking styles in the journal entries.

The Linguistic Inquiry and Word Count (LIWC) analysis in Table 6.3 reveals nuanced differences in how participants engage with contextual and generic journaling prompts. Notably, generic prompts elicit slightly longer responses (Mean = 44.51, SD = 15.78) compared to contextual prompts (Mean = 43.67, SD = 17.62). This difference in response length may be related to the thinking style encouraged by
Table 6.3: LIWC Scores: Below we present a selection of representative LIWC categories, along with their mean and standard deviation (SD) values, categorized by Contextual and Generic Journals.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Contextual Journals</th>
<th>Generic Journals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Word count</td>
<td>43.67</td>
<td>17.62</td>
</tr>
<tr>
<td>Analytical thinking</td>
<td>33.39</td>
<td>27.34</td>
</tr>
<tr>
<td>Clout</td>
<td>10.78</td>
<td>13.23</td>
</tr>
<tr>
<td>Authenticity</td>
<td>83.78</td>
<td>18.76</td>
</tr>
<tr>
<td>Emotional Tone</td>
<td>67.49</td>
<td>28.45</td>
</tr>
<tr>
<td>Affect</td>
<td>5.84</td>
<td>4.36</td>
</tr>
<tr>
<td>Positive tone</td>
<td>4.67</td>
<td>3.82</td>
</tr>
<tr>
<td>Negative tone</td>
<td>1.03</td>
<td>1.70</td>
</tr>
<tr>
<td>Pronouns</td>
<td>18.14</td>
<td>5.95</td>
</tr>
<tr>
<td>Cognition</td>
<td>14.69</td>
<td>7.26</td>
</tr>
<tr>
<td>Insight</td>
<td>3.43</td>
<td>3.12</td>
</tr>
<tr>
<td>Drives</td>
<td>4.67</td>
<td>3.82</td>
</tr>
<tr>
<td>Social processes</td>
<td>7.95</td>
<td>6.37</td>
</tr>
<tr>
<td>Social behavior</td>
<td>4.00</td>
<td>3.90</td>
</tr>
<tr>
<td>Social referents</td>
<td>3.34</td>
<td>3.75</td>
</tr>
<tr>
<td>Time orientation</td>
<td>5.63</td>
<td>4.25</td>
</tr>
<tr>
<td>Past focus</td>
<td>5.17</td>
<td>4.62</td>
</tr>
<tr>
<td>Present focus</td>
<td>5.53</td>
<td>4.30</td>
</tr>
<tr>
<td>Future focus</td>
<td>1.38</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Each prompt type. Generic prompts yield higher analytic thinking scores (Mean = 40.50, SD = 21.72), indicating a more formal and logical thinking style. In contrast, contextual prompts result in lower analytic thinking scores (Mean = 33.39, SD = 27.34), suggesting a more personal and spontaneous writing approach. This difference in thinking styles is also reflected in the Clout scores, which reveal a disparity in the level of confidence and expertise conveyed through language. Contextual journal entries express less confidence and authority (Mean = 10.78, SD = 13.23) compared to generic journals (Mean = 9.43, SD = 9.53), resulting in a more tentative and exploratory writing tone. Furthermore, the two prompt types also differ in terms of authenticity, emotional tone, and cognitive processes. Contextual journals score lower in authenticity (Mean = 83.78, SD = 18.76) compared to generic journals (Mean = 85.11, SD = 20.94), but show similar emotional tone scores (contextual Mean = 67.49, SD = 28.45; generic Mean = 68.09, SD = 25.41). But, generic prompts prompted a
noticeably higher affective content (Mean = 8.17, SD = 3.80) compared to contextual prompts (Mean = 5.84, SD = 4.36), suggesting that generic prompts may encourage broader emotional expressions. In addition, generic prompts have a higher positive tone (e.g., good, well, new, love) and reduced negative tone (e.g., bad, wrong, too much, hate) as well. However, contextual journals have higher cognition scores (Mean = 14.69, SD = 7.26) than generic journals (Mean = 11.41, SD = 6.44), indicating a greater emphasis on thinking, problem-solving, and memory recall. Moreover, the language used in contextual journals reveals a greater focus on personal experiences and relationships, with more pronouns used (Mean = 18.14, SD = 5.95) compared to generic journals (Mean = 14.66, SD = 5.02). This is consistent with the finding that contextual prompts encourage more social references (Mean = 3.34, SD = 3.75) than generic journals (Mean = 1.87, SD = 2.09), indicating a strong focus on social bonds and community. Finally, generic prompts encourage a broader temporal focus (Mean = 7.38, SD = 4.41) compared to contextual prompts (Mean = 5.63, SD = 4.25), particularly in the higher scores for past and future focus. This suggests that generic prompts may encourage participants to link their current experiences with past memories or future aspirations more frequently than contextual prompts. Note that due to the differing time periods associated with each type of journaling, we normalized the scores to ensure comparability. To do this, we first calculated the LIWC scores per week for each participant during the 6-week contextual prompts period and the 2-week generic prompts period, separately. Then, we averaged each set of scores separately to obtain a final weekly average for both journaling experiences.

Self-Compassion and Gratitude Inducing Journaling: Before starting their journal entries, we ask participants to indicate their mood using an emoji scale from 1 to 5, ranging from sad to happy. Out of the 533 contextual journaling responses we receive, participants rate their mood with a score of 1 on 5 occasions, 2 on 28 occasions, and
3 on 109 occasions. They most frequently report feeling better, with scores of 4 and 5 occurring 306 and 85 times, respectively. As part of our journal prompt generation strategy, whenever participants indicate not feeling good (characterized by a score of 3 or below, which happens in approximately 27% of instances), we offer a journaling prompt aimed at fostering gratitude or self-compassion through the MindScape app.

The objective of these tailored prompts is to transform any negative emotions into a more positive, affirming experience. To evaluate the journal entries received for these prompts, we conduct a two-stage analysis process: firstly, we leverage GPT-4 to analyze the presence of gratitude or self-compassion in these journaling entries, clarifying whether the journal prompts achieve their intended effect of fostering positive emotional responses. Next, to gain a more nuanced understanding of the emotions associated with the journal responses, we utilize a state-of-the-art emotion recognition model. This model, trained on the ‘go_emotions’ dataset with a RoBERTa architecture, is designed for multi-label classification. RoBERTa is a transformers model pretrained on a large corpus of English data in a self-supervised fashion and the ‘go_emotions’ dataset is derived from Reddit data and includes 28 emotion labels. It is a multi-label dataset where one or multiple labels may apply to any given input text. Consequently, this model provides 28 probability outputs (i.e., representing 28 emotion labels) for each input, reflecting the complex emotional dimensions captured in the text.

Out of 142 journal entries submitted by participants, where they rated their feelings as less than 3 on the emoji scale, GPT-4 identifies the presence of gratitude or self-compassion in 48 entries, representing 34% of the journals. It is important to note that this analysis focuses on detecting expressions of gratitude or self-compassion, rather than examining the overall sentiment of the text (i.e., whether positive or negative). To further analyze the emotional content, we employ an emotion recognition
6.4 Results Contextual AI Journaling

(a) Gratitude/Self-Compassion Detected  
(b) Gratitude/Self-Compassion Not Detected

Figure 6.8: Most Frequently Occurring Emotions: The two word clouds displayed above illustrate the most commonly encountered emotions in journal entries responding to contextual prompts designed to elicit self-gratitude and self-compassion. Figure (a) represents emotions from journal entries where GPT-4 detected the presence of gratitude or self-compassion, while Figure (b) shows emotions from entries where GPT-4 did not detect these sentiments. Although both word clouds feature similar words, negative emotions such as disappointment, disapproval, annoyance, sadness, and anger are more prominent in Figure (b).

model to label the emotions associated with each text. We select the top 10 labels from each and create word clouds to visualize the most frequent emotions detected in these journals. In Figure 6.8a, we present the word cloud for texts where GPT-4 detects gratitude or self-compassion. In contrast, Figure 6.8b displays the word cloud for the remaining 66% of texts where the LLM does not detect either gratitude or self-compassion. Both word clouds show largely similar emotion labels, but with some notable differences. Emotion labels like disappointment, disapproval, annoyance, sadness, and anger appear with lower frequency in the gratitude cloud, while they are more pronounced in the cloud where gratitude and self-compassion are not detected by the LLM. Despite this, we observe that most of the journals participants submit when feeling low predominantly feature emotions like approval, realization, neutrality, and optimism. This suggests that even in cases where we cannot detect explicit expressions of self-compassion or gratitude, the reflections might still be steered towards
positive outcomes. Please refer to Appendix C.6 for examples of journals focusing on gratitude and self-compassion.

6.4.4. Exploring Changes in Wellbeing and Emotional Growth

In this section, we evaluate the changes in the participants’ well-being and personal growth following the contextual journaling phase. We administer several standardized surveys to participants at multiple stages of the study: baseline, weekly intervals, and follow-up. These surveys are designed to assess changes in their behavior and well-being.

**Changes in Baseline vs. Follow-up Survey:** We administer the same set of standard surveys to participants at the beginning of the study and at the six-week follow-up, when the contextual journaling phase ends and generic journaling begins. We compare the responses and detail the mean differences on Table 6.4, which includes the baseline mean (start of the study), mean at follow-up (at the six-week mark), the mean change in value, mean change expressed as a percentage, the effect size (Cohen’s d), and the 95% confidence interval (C.I.) of the effect size. We also conduct a paired t-test, shading non-statistically significant values in grey. It is essential to note that, while we report statistical significance in adherence to standard result reporting practices, considering outcomes regardless of statistical significance is valuable given our small sample size. This limitation often leads to fewer statistically significant results. Therefore, we also report effect sizes, which reveal notable effects despite a lack of significance. Given the exploratory nature of our study, dismissing potential relationships solely based on statistical significance is not advisable. Moreover, the wide confidence intervals for all values—attributable to the small sample size—present intriguing results that warrant further investigation with a larger sample in future studies.

We observe several positive outcomes from the study. Interestingly, we find a
Table 6.4: Changes at Follow-Up: We assess changes in follow-up surveys relative to the baseline. We calculate the effect size using Cohen’s D, and “C.I.” denotes the 95% confidence interval for the effect size. We perform paired t-tests, and we highlight statistically insignificant results in gray. (** p-value ≤ .01, * .05 < p-value ≤ .10).

<table>
<thead>
<tr>
<th>Facet</th>
<th>Baseline</th>
<th>Follow-up</th>
<th>Mean Change</th>
<th>Percentage Change</th>
<th>Effect Size</th>
<th>C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality (260)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extraversion</td>
<td>5.15</td>
<td>5.25</td>
<td>0.10</td>
<td>1.94%</td>
<td>0.08</td>
<td>(-0.39, 0.54)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>6.60</td>
<td>6.80</td>
<td>0.20</td>
<td>3.03%</td>
<td>0.21</td>
<td>(-0.25, 0.67)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>7.15</td>
<td>7.35</td>
<td>0.20</td>
<td>2.80%</td>
<td>0.17</td>
<td>(-0.30, 0.63)</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>7.20</td>
<td>6.35</td>
<td>-0.85</td>
<td>-11.81%</td>
<td>-0.63</td>
<td>(-1.09, -0.16)</td>
</tr>
<tr>
<td>Openness</td>
<td>6.90</td>
<td>6.90</td>
<td>0.00</td>
<td>0.00%</td>
<td>0.00</td>
<td>(-0.46, 0.46)</td>
</tr>
<tr>
<td>Emotion Regulation (256)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive reappraisal</td>
<td>12.90</td>
<td>13.25</td>
<td>0.35</td>
<td>2.71%</td>
<td>0.07</td>
<td>(-0.39, 0.54)</td>
</tr>
<tr>
<td>Expressive suppression</td>
<td>11.65</td>
<td>12.40</td>
<td>0.75</td>
<td>6.44%</td>
<td>0.15</td>
<td>(-0.32, 0.61)</td>
</tr>
<tr>
<td>Affect (337)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive affect</td>
<td>31.45</td>
<td>33.70</td>
<td>2.25*</td>
<td>7.15%</td>
<td>0.39</td>
<td>(-0.07, 0.86)</td>
</tr>
<tr>
<td>Negative affect</td>
<td>25.00</td>
<td>22.35</td>
<td>-2.65**</td>
<td>-10.60%</td>
<td>-0.62</td>
<td>(-1.08, -0.15)</td>
</tr>
<tr>
<td>Stress &amp; Anxiety</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived stress (69)</td>
<td>7.60</td>
<td>6.75</td>
<td>-0.85</td>
<td>-11.18%</td>
<td>-0.28</td>
<td>(-0.74, 0.19)</td>
</tr>
<tr>
<td>State-trait anxiety (200)</td>
<td>45.83</td>
<td>42.17</td>
<td>-3.67</td>
<td>-8.00%</td>
<td>-0.30</td>
<td>(-0.77, 0.16)</td>
</tr>
<tr>
<td>Resilience (294)</td>
<td>2.91</td>
<td>2.94</td>
<td>0.03</td>
<td>1.03%</td>
<td>0.10</td>
<td>(-0.36, 0.57)</td>
</tr>
<tr>
<td>Psychological Wellbeing (269)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomy</td>
<td>8.05</td>
<td>7.70</td>
<td>-0.35</td>
<td>-4.35%</td>
<td>-0.16</td>
<td>(-0.62, 0.31)</td>
</tr>
<tr>
<td>Personal growth</td>
<td>5.45</td>
<td>5.50</td>
<td>0.05</td>
<td>0.92%</td>
<td>0.03</td>
<td>(-0.43, 0.49)</td>
</tr>
<tr>
<td>Positive relations</td>
<td>7.00</td>
<td>6.10</td>
<td>-0.90**</td>
<td>-12.86%</td>
<td>-0.48</td>
<td>(-0.95, -0.01)</td>
</tr>
<tr>
<td>Purpose</td>
<td>5.50</td>
<td>5.75</td>
<td>0.25</td>
<td>4.55%</td>
<td>0.22</td>
<td>(-0.24, 0.69)</td>
</tr>
<tr>
<td>Self-acceptance</td>
<td>6.95</td>
<td>6.60</td>
<td>-0.35</td>
<td>-5.04%</td>
<td>-0.19</td>
<td>(-0.66, 0.27)</td>
</tr>
<tr>
<td>Life satisfaction (89)</td>
<td>22.35</td>
<td>22.70</td>
<td>0.35</td>
<td>1.57%</td>
<td>0.09</td>
<td>(-0.38, 0.55)</td>
</tr>
<tr>
<td>Flourishing (90)</td>
<td>42.60</td>
<td>43.45</td>
<td>0.85</td>
<td>2.00%</td>
<td>0.14</td>
<td>(-0.32, 0.61)</td>
</tr>
<tr>
<td>Social and Interpersonal Wellbeing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social provision (59)</td>
<td>16.80</td>
<td>16.25</td>
<td>-0.55</td>
<td>-3.27%</td>
<td>-0.18</td>
<td>(-0.64, 0.29)</td>
</tr>
<tr>
<td>Loneliness</td>
<td>8.50</td>
<td>7.95</td>
<td>-0.55*</td>
<td>-6.47%</td>
<td>-0.42</td>
<td>(-0.88, 0.05)</td>
</tr>
<tr>
<td>Cognition and Self-Awareness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mindfulness (20)</td>
<td>44.35</td>
<td>47.35</td>
<td>3.00**</td>
<td>6.76%</td>
<td>0.55</td>
<td>(0.07, 1.01)</td>
</tr>
<tr>
<td>Self-reflection (292)</td>
<td>29.30</td>
<td>31.00</td>
<td>1.70**</td>
<td>5.80%</td>
<td>0.47</td>
<td>(0.00, 0.93)</td>
</tr>
<tr>
<td>Insight (292)</td>
<td>25.75</td>
<td>27.70</td>
<td>1.95*</td>
<td>7.57%</td>
<td>0.36</td>
<td>(-0.10, 0.82)</td>
</tr>
</tbody>
</table>

significant decrease of 11.81% in the personality trait neuroticism, which is typically associated with negative emotions, with a medium effect size (p-value = 0.001, effect size (d) = -0.63). We consider effect sizes of 0.2, 0.5, and 0.8 as small, medium, and large, respectively, regardless of the sign, which merely indicates the direction of
change. Although changes in agreeableness are not statistically significant, we observe a modest increase of 3.03% with a small effect size. Changes in other personality traits do not reach statistical significance, and their effect sizes remain small. Given that personality is generally stable, significant alterations in traits like neuroticism and agreeableness within a short timeframe are noteworthy. We do not observe any significant changes in emotion regulation, neither through statistical significance nor through effect sizes. However, we note promising indicators of improved well-being at the follow-up, including an increase in positive affect and a decrease in negative affect. Specifically, positive affect, which reflects the extent to which individuals experience positive moods such as joy, interest, and alertness, increases by 7.15% \((p-value = 0.05, d = 0.39)\). Conversely, negative affect, which encompasses a range of negative emotional states including anxiety, depression, stress, and sadness, decreases by 10.60% \((p-value = 0.05, d = -0.62)\). Both changes are statistically significant and exhibit moderate to large effect sizes.

We observe notable changes in various psychological metrics. Stress and anxiety decrease by 11.18% and 8.00%, respectively, although these reductions are not statistically significant and exhibit small effect sizes. Resilience increases by 1.03%, but this change is not statistically significant and demonstrates a very low effect size. In terms of psychological well-being, the results are mixed. Autonomy, defined as being self-determining and independent, decreases by 4.35% \((p-value = 0.49, d = -0.16)\). Positive relations with others, which encompass warm, satisfying, trusting relationships, decrease by 12.86% \((p-value = 0.04, d = -0.48)\), and self-acceptance, referring to a positive attitude toward oneself, decreases by 5.04% \((p-value = 0.40, d = -0.19)\). Only the decrease in positive relations is statistically significant and exhibits a medium effect size. However, other elements within psychological well-being show positive changes. We observe a 0.92% increase in personal growth \((p-value = \)
0.89, \( d = 0.03 \)), which involves seeing improvement in oneself and behavior over time. Purpose in life, defined as having goals in life and a sense of directedness, increases by 4.55\% (\( p\text{-value} = 0.32, \ d = 0.22 \)). Life satisfaction, an evaluation of a person’s quality of life, increases by 1.57\% (\( p\text{-value} = 0.69, \ d = 0.09 \)). Flourishing—self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism—increases by 2.00\% (\( p\text{-value} = 0.53, \ d = 0.14 \)). Although these results are statistically insignificant and associated with small effect sizes, they indicate promising trends. Additionally, we find a statistically insignificant decrease of 3.27\% in social provision, specifically perceived social support, with a very small effect size (\( p\text{-value} = 0.44, \ d = -0.18 \)). On the other hand, subjective feelings of loneliness show statistically significant improvement, decreasing by 6.47\% with a medium effect size (\( p\text{-value} = 0.07, \ d = -0.42 \)).

We observe exclusively positive outcomes in cognition and self-awareness. Mindfulness, self-reflection—defined as the inspection and evaluation of one’s thoughts, feelings, and behaviors—and insight, which refers to a clear understanding of one’s mental and emotional processes, all show significant increases. Each of these dimensions demonstrates statistically significant improvements with small to medium effect sizes. Specifically, we observe a 6.76\% surge in mindfulness with a medium effect size (\( p\text{-value} = 0.02, \ d = 0.55 \)); self-reflection rises by 5.80\% with a medium effect size (\( p\text{-value} = 0.04, \ d = 0.47 \)); and insight grows by 7.57\% with a small effect size (\( p\text{-value} = 0.10, \ d = 0.36 \)). These results indicate that the contextual journaling integral to the study substantially enhances the key factors we aimed to influence: self-awareness, self-monitoring, and clarity of self-perception.

**Weekly Changes:** We employ the MindScape app to administer weekly ecological momentary assessments (EMA) to participants. These assessments comprise
the Patient Health Questionnaire-4 (PHQ4) (168), Self-reflection and Insight Scale (SRIS) (292), 5-item Mindful Attention Awareness Scale (MAAS) (54), and the 10-item Positive and Negative Affect Schedule (PANAS) (310). Every Sunday, the app sends notifications to participants, prompting them to complete the surveys. We utilize a mixed-effects model to examine changes in participants’ scores over the weeks, accounting for their self-reported gender, student status (graduate or undergraduate), past journaling experience, and race (‘multiple’ race category is merged into ‘other’ for simplicity). The mixed-effects model we apply to analyze the outcome scores is formulated as follows:

\[
\text{outcome}_i = \beta_0 + \beta_1 \text{week}_i + \beta_2 \text{gender\_Male}_i + \beta_3 \text{gender\_Nonbinary}_i \\
+ \beta_4 \text{race\_Black}_i + \beta_5 \text{race\_Other}_i + \beta_6 \text{race\_White}_i + \beta_7 \text{journaling\_exp}_i \\
+ \beta_8 \text{student\_status}_i + b_{0i} + b_{1i} \text{week}_i + \epsilon_{ij}
\] (6.1)

where:

- \(\text{outcome}_i\) refers to the scores obtained from the \{PHQ4, SRIS, MAAS, PANAS\} surveys for the \(i\)-th subject at the \(j\)-th week.
- \(\beta_0, \beta_1, \ldots, \beta_8\) are the fixed coefficients for intercept, week, and other covariates.
- \(b_{0i}\) is the random intercept for the \(i\)-th subject.
- \(b_{1i}\) is the random slope for the \(i\)-th subject associated with the effect of week.
- \(\epsilon_{ij}\) is the residual error.

We compute the outcomes by first utilizing the total PHQ4 score as-is, and then deriving two subscores: anxiety, which is the sum of the first two items, and depression, which is the sum of the last two items. Additionally, we generate scores
6.4 Results

for positive affect and negative affect from the PANAS, and calculate scores for self-reflection and insight from the SRIS. Furthermore, we compute a total mindfulness score from the MAAS. We then incorporate all these variables as outcomes in the mixed-effects model, examining their changes and relationships over time. By doing so, we can gain a comprehensive understanding of how the participants’ mental health and well-being evolve throughout the study.

Table 6.5: Weekly EMA Changes: We analyze the changes in self-reported EMA responses throughout the study period using a mixed-effects model. The number outside the brackets represents the model’s coefficient, while the number inside the brackets indicates the standard error. We highlight statistically insignificant results in gray. (**p-value ≤ .01, **.01 < p-value ≤ .05, *.05 < p-value ≤ .10).

<table>
<thead>
<tr>
<th>EMA</th>
<th>Week</th>
<th>Gender Male</th>
<th>Gender Nonbinary</th>
<th>Race White</th>
<th>Race Black</th>
<th>Race Other</th>
<th>Journaling Experience</th>
<th>Student Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ4 (168)</td>
<td>-0.25 (0.08)**</td>
<td>-1.74 (0.98)*</td>
<td>-1.85 (2.54)</td>
<td>0.72 (1.35)</td>
<td>-0.12 (1.35)</td>
<td>1.46 (1.29)</td>
<td>-0.98 (1.46)</td>
<td>-1.74 (1.22)</td>
</tr>
<tr>
<td>Anxiety (168)</td>
<td>-0.12 (0.05)**</td>
<td>-1.05 (0.54)*</td>
<td>-1.18 (1.42)</td>
<td>0.93 (0.76)</td>
<td>0.34 (0.75)</td>
<td>0.71 (0.72)</td>
<td>0.04 (0.81)</td>
<td>-0.18 (0.68)</td>
</tr>
<tr>
<td>Depression (168)</td>
<td>-0.13 (0.05)***</td>
<td>-0.69 (0.52)</td>
<td>-0.66 (1.37)</td>
<td>-0.19 (0.73)</td>
<td>-0.44 (0.72)</td>
<td>0.74 (0.69)</td>
<td>-1.01 (0.78)</td>
<td>-1.55 (0.65)***</td>
</tr>
<tr>
<td>Positive affect (310)</td>
<td>0.04 (0.10)</td>
<td>1.67 (1.10)</td>
<td>2.31 (2.87)</td>
<td>-0.61 (1.53)</td>
<td>1.51 (1.52)</td>
<td>2.15 (1.44)</td>
<td>0.69 (1.64)</td>
<td>2.45 (1.37)*</td>
</tr>
<tr>
<td>Negative affect (310)</td>
<td>-0.15 (0.12)</td>
<td>-2.03 (1.52)</td>
<td>-2.60 (3.93)</td>
<td>-0.39 (2.10)</td>
<td>-0.36 (2.09)</td>
<td>0.10 (1.99)</td>
<td>-2.50 (2.26)</td>
<td>-2.22 (1.89)</td>
</tr>
<tr>
<td>Mindfulness (54)</td>
<td>0.04 (0.04)</td>
<td>0.84 (0.61)</td>
<td>2.20 (1.58)</td>
<td>0.72 (0.85)</td>
<td>0.78 (0.84)</td>
<td>-0.49 (0.81)</td>
<td>1.44 (0.91)*</td>
<td>1.31 (0.76)*</td>
</tr>
<tr>
<td>Self-reflection (292)</td>
<td>0.39 (0.14)**</td>
<td>2.02 (3.22)</td>
<td>7.56 (8.23)</td>
<td>10.92 (4.42)</td>
<td>-1.31 (4.42)</td>
<td>3.33 (4.24)***</td>
<td>14.21 (4.78)***</td>
<td>2.34 (4.02)</td>
</tr>
<tr>
<td>Insight (292)</td>
<td>0.04 (0.15)</td>
<td>3.76 (4.15)</td>
<td>12.56 (10.56)</td>
<td>1.64 (5.68)</td>
<td>2.66 (5.68)</td>
<td>-3.36 (5.46)</td>
<td>3.22 (6.14)</td>
<td>6.13 (5.18)</td>
</tr>
</tbody>
</table>

We display the results in Table 6.5, which reveals several remarkable findings. Note that the number outside the brackets represents the coefficients, while the number inside the brackets indicates the standard error. Notably, anxiety levels consistently decrease each week, with a statistically significant reduction (p-value = 0.01, β = -0.12). Interestingly, this decrease is more pronounced in males (p-value = 0.06, β = -1.05). Our analysis also shows a significant decrease in depression scores over time (p-value = 0.01, β = -0.131), particularly among graduate students compared to undergraduates (p-value = 0.01, β = -1.55). This suggests that journaling may be more effective for graduate students, although no other demographic factors like gender show significant effects. The overall PHQ4 score also demonstrates a decreasing trend during the study, more so in males. In terms of affective states, we observe no significant changes in negative affect over the study period. However, graduate
students report higher levels of positive affect compared to undergraduate students. This stability in affect contrasts with findings from the baseline and follow-up surveys reported in Section 6.4.4, where we observe a statistically significant increase in positive affect and a decrease in negative affect. Similarly, MindScape boosts self-reflective capacities, as evident from a significant weekly increase in self-reflection scores ($p-value = 0.01$, $\beta = 0.39$). Although gender differences in self-reflection are not statistically significant, we observe a notable trend suggesting that race may influence outcomes. Participants with prior journaling experience benefit more ($p-value = 0.01$, $\beta = 14.23$), showing greater score in self-reflection. This aligns with our earlier comparison of baseline and follow-up scores in self-reflection, which yielded positive and statistically significant results. In contrast, insights and mindfulness, measured by their respective total scores, do not exhibit significant changes throughout the study. However, participants with prior journaling experience and graduate students experience enhanced mindfulness benefits compared to others ($p-value = 0.10$, $\beta = 1.44$).

(a) Increase in Self-reflection EMA Score 
(b) Decrease in PHQ4 EMA Score

Figure 6.9: Improvement in Well-Being as Indicated by EMAs: Both figures demonstrate improvements in well-being, either through an increase in self-reflection (Figure a) or a decrease in PHQ-4 scores, which measure depression and anxiety (Figure b). The numbers in the boxes at the bottom indicate the count of unique participants who submit the survey in each respective week.

We illustrate the increase in self-reflection scores and the decrease in PHQ-4 scores
in Figures 6.9a and 6.9b, respectively. The numbers at the bottom of the figures in boxes represent the number of participants from whom we receive EMA responses in each specific week. On average, we receive responses from 15 participants over the course of the first six weeks. However, in weeks 7 and 8, which mark the beginning of the generic journaling period, the numbers drop to 6 and 3, respectively. The shaded areas represent the 95% confidence intervals. Overall, it appears that participants experience several positive changes during the study period. These improvements across various psychological dimensions, particularly in cognition and self-awareness, demonstrate the efficacy of contextual journaling. While some areas show minimal changes or declines, the significant positive trends may indicate the potential beneficial impact of the study on enhancing participants’ mental well-being and self-related cognitions. This overall positive shift indicates promising paths for future applications and studies aimed at further understanding and supporting mental health and cognitive awareness through contextual journaling.

6.4.5. MindScape App Performance

Upon concluding the study, we solicit feedback from participants regarding the app’s performance, their experience with contextual journaling, and related topics. We present our findings from the analysis of these responses below.

First, we administer the standard System Usability Scale (SUS) survey to participants. The SUS is a widely used survey comprising 10 questions designed to assess the perceived usability of a system. In this instance, it evaluates the MindScape app. Figure 6.10 displays the participant responses, including a dotted line indicating the typical thresholds for poor, good, and excellent SUS scores. Our analysis reveals that 50% (N=10) of participants rate the system as excellent, with a score of 80 or higher. Additionally, 35% (N=7) rate the system as good (a score of 68 or higher), while the remaining 15% (N=3) rate the system as poor, scoring below 68.
6.4 Results Contextual AI Journaling

Figure 6.10: System Usability Score: MindScape receives high usability ratings. In the figure, the red dotted line marks the threshold for poor usability, the blue dotted line for good usability, and the green dotted line for excellent usability. Over 80% of participants rate the MindScape system as either good or excellent.

We then present participants with several other questions on their experience with the app. We discuss some responses here and visualize these responses in Figure 6.11. Please see Appendix C.1 for a complete list of questions and participant responses.

We ask about the relevance of the journaling prompts generated by the MindScape app and find that 10% (N=2) consider them very relevant, 60% (N=12) moderately relevant, and 30% (N=6) slightly relevant. When assessing the check-in prompts specifically, 5% (N=1) indicate they are not at all relevant, 40% (N=8) find them slightly relevant, 40% (N=8) think they are moderately relevant, and 15% (N=3) consider them very relevant.

To enrich our understanding of participants’ experiences using the app, we supplement our quantitative analysis with qualitative, open-ended questions. We ask participants if they recall any specific prompts from the MindScape app that signif-
6.4 Results

Contextual AI Journaling

icantly resonated with them or were particularly relevant to their experiences. This inquiry aims to uncover how the app’s context-aware prompts facilitate deeper self-understanding or self-awareness. Participants report various responses that illustrate the personalized impact of these prompts. Many highlight how specific questions related to their daily routines or habits prompt meaningful reflection. For instance, one participant mentions a prompt about their walking routine, which coincides with a new meditative practice: “About a week or two after I started it, I got a question from the app about my walking routine, and it gave me an opportunity to reflect on how the meditative practice had been going!” This example underscores how timely and relevant prompts enhance mindfulness and self-awareness. Other participants appreciate prompts that encourage proactive planning and goal-setting, such as those urging them to think about weekly exercise goals. This not only makes them contemplate their physical health but also motivates them to set concrete plans. Additionally, reminders about social interactions lead some users to reach out to friends more frequently, improving their social life. One user points out, “A lot of the prompts remind me that I have not been socializing in person as much, which has led to me reaching out to friends I do not see as often.” The app’s ability to track changes in routines and lifestyle also stands out for many users. For example, prompts related to changes in workspaces or sleep patterns provide insights that participants might not have noticed on their own. One participant succinctly notes, “The app has been accurately keeping track of my changes in work routines, which allows me to take time and reflect on how these changes are affecting my overall work performance.”

Given that these questions are posed at the end of the study, following two weeks of generic prompts, we ask participants to self-report on the frequency with which context-aware prompts facilitate more profound reflection than usual. Out of respondents, 5% (N=1) note that it never leads to deeper reflections, 10% (N=2) say it
Figure 6.11: Participants’ Experience with MindScape: This figure visualizes responses to various questions about participants’ experiences with the app. Each person icon represents one participant, with icons and their corresponding responses color-coded to match specific response labels.

rarely does, while the majority report that it sometimes (45%, N=9), often (30%, N=6), or always (10%, N=2) leads to more in-depth reflection.

Fifteen percent of participants report moderate changes in their daily habits or behaviors since using the MindScape app, while 30% notice slight changes. However, 55% do not report seeing any changes. Despite this, when we ask participants to share specific instances where they noticed a change in their behavior or habits due to using the app, many highlight tangible changes driven by the app. For instance, one user notes the impact of visual prompts related to phone usage, stating, “Every time I see that little green light [indicating microphone use], I’m prompted to think about whether my current device usage aligns with my goals.” This increased awareness often leads them to decrease unreflective screen time and transition to other daily activities. Another participant describes seamlessly integrating the journaling habit into their daily routine, aligning it with brushing their teeth each night.
6.4 Results

Contextual AI Journaling

This structured approach helps stabilize the practice as part of their existing routine. Similarly, changes in social interactions are notable, with one respondent mentioning significant improvements: “I’ve seen improvement in my social life...through prompts that make me reflect on the importance of nurturing conversations with my friends.” Additionally, the app enhances self-awareness regarding personal well-being and daily activities. Reports include better monitoring of sleep patterns, more frequent walks, and increased engagement in self-reflection and meditation. However, not all feedback indicates a change. One participant remarks on the challenging nature of adapting behaviors during a particularly busy life phase, suggesting that while the app identifies reduced phone usage, the decrease is more due to their hectic schedule rather than a conscious effort spurred by the app.

Recognizing that behavioral changes can be challenging for various reasons, we included an additional question to gather more insight: “Has using the MindScape app influenced the way you plan or structure your week? If so, how? Additionally, has the process of journaling and reflecting with the app altered your mindset, leading you to appreciate your daily activities more? Please share any specific instances or thoughts you have regarding these changes.” Participants’ responses reveal varied impacts of the MindScape app on their daily planning and mindset. Several appreciate the structuring aspect introduced by the app, notably through timed journaling activities. Others value the simplicity and reminder functions of the app, which facilitate engagement in desired activities. One participant notes, “It’s led me to prioritize working out and socializing in person more...the prompts are good reminders.” The app significantly impacts mindfulness and daily appreciation for some users. One says, “It has made me more aware of small positive interactions I’ve had throughout the day...the app brought those moments closer again.” Another remarks on lifestyle improvements, “I’ve been trying to have better sleep habits, enjoy the outdoor world, and
the app’s feedback on activities like walks is nice.” However, some find the prompts and check-ins inadequate or misaligned with their personal reflection needs. For instance, one user points out the incongruence in the app’s social interaction tracking, stating, “The check-ins seem to assume that whenever I am talking, it’s social, but I work and go to school and have kids, so none of this is what I consider social.”

Moving on, 55% (N=11) of participants express satisfaction or high satisfaction with the app, while 30% (N=6) remain neutral, and 15% (N=3) feel dissatisfied. The ease of integrating the MindScape app into daily routines is also notable, with 75% (N=15) stating it is somewhat easy to very easy. Regarding comfort with using data from phones and smartwatches alongside AI to personalize journaling prompts, 70% (N=14) voice comfort or great comfort with the notion, while 30% (N=6) are uncomfortable or neutral. In terms of data privacy and security, 60% (N=12) express no or slight concern, whereas the remaining 40% (N=8) report moderate to high levels of concern. When asked about differences in the depth and quality of their reflections when using context-aware directed prompts from the MindScape app compared to standard journaling, several participants appreciate the structured prompts for enhancing their engagement and reflection practices. For instance, one user mentions, “With MindScape, I find journaling more fun and interactive,” pointing out that the app eliminates the often cumbersome task of recollecting daily events. This sentiment is echoed by another participant who notes, “Previously, it was cumbersome for me to journal as I’d have to sit and recollect everything. With MindScape, I reflect more earnestly and compare my current state with previous ones, giving me clear insights into my mental and physical well-being.” Participants also value the specific insights prompted by the app, which sometimes bring attention to overlooked aspects of their daily lives. As one participant shares, “The context-aware prompts pull out surprising trends that I may not have noticed and ask me to reflect on it.” This feature helps
some users gain a deeper understanding of their behavioral patterns and encourages proactive thinking, like another user who points out, “The prompts lead me to think about specific parts of the experiences of my day and how I might make changes.” However, challenges with the specificity and relevance of the AI-driven prompts are also noted. Some participants feel that the prompts can be too rigid or not entirely reflective of their true daily experiences due to inaccuracies in activity or location tracking. For example, one user criticizes the prompts for being “sometimes based on metrics that might be irrelevant,” such as reacting to an unusually high number of spam calls as if they were meaningful phone conversations. Similarly, another points out that “the app tends to see trends or differences in my behavior when there isn’t a particular cause behind them.” Despite these challenges, the general consensus acknowledges the utility of the MindScape app in fostering regular reflective practices. Even those who note limitations often recognize the benefits of prompted reflections, with remarks like “I think that I can more easily decide what to write about with a prompt, but I don’t know if I access my feelings in the same way as a free-form journal.”

6.4.6. Strengths, Weaknesses and Improvement Opportunities

At the conclusion of the study, we gather qualitative, open-ended responses from participants regarding the app’s strengths and weaknesses. We synthesize their responses in this section.

The MindScape app receives positive feedback from users for its innovative approach to integrating behavioral patterns with journal prompts. Participants recognize and value the regular notifications as powerful tools for maintaining daily reflection habits. For example, a participant notes, “I liked how it regularly made me aware of what I am doing, and it helped me reflect on the activities. I loved the journal prompts as they were not too specific nor broad that I could sufficiently elab-
orate on my thoughts.” Another user reflects, “It did initiate some reflection on the interactions I had throughout the day, how I value the people in my life”. Users also note the app’s ease of use as a major advantage, making it accessible and straightforward for daily interactions. “Love the ease of use of the app. Super simple to click on the notification and complete the journaling and check-in practices,” as one user expresses. Another adds, “I also enjoyed the personalized prompts and how I only had to write short pieces for the journal.”

Participants also identify several areas for improvement crucial for refining the app’s functionality. The most common comment from participants is that the accuracy of the context-aware prompts needs enhancement. Users report instances where the app does not accurately recognize their location or current activity, highlighting the need for more precise behavioral sensing. “Many times, the app didn’t seem to be aware of where I lived and would ask me questions about being in my dorm when I wasn’t there,” a participant shares, suggesting that increased accuracy could significantly improve user experience. Achieving precise context sensing with passive data is challenging, but integrating additional data sources, such as smartwatches, could help address some of this issue. In fact, a few participants suggest they want the ability to link the app to their fitness trackers. Related to that, the repetitiveness of the prompts also emerges as a concern, which could potentially be improved with the help of more signals such as those from smartwatches. A few users note receiving identical prompts over multiple days, diminishing the personalized experience the app aims to provide. Ensuring more dynamic and varied prompt generation could sustain user interest and engagement. Additionally, a few participants raise concerns about the app’s battery consumption, indicating a need for optimizing power management to enhance overall usability.

We highlight some improvement opportunities based on participant feedback be-
(a) **Enhanced Personalization and Prompt Variability:** Participants suggest enhancing personalization by considering responses to previous prompts when generating new ones. The app currently uses behavioral data to create prompts, but incorporating responses from prior prompts could introduce greater diversity and relevance to the questions posed. Participants prefer prompts that explore personality aspects or broader life reflections beyond mere lifestyle habits. While we introduce broader questions on weekends, feedback suggests refining this balance. By analyzing users’ previous responses, the app can better determine whether to generate prompts based on sensed data or earlier interactions. One user commented, “If previous prompt responses influence future prompts more, that could help keep somewhat of a narrative going between journal reflections.” Additionally, some participants indicate that more open-ended questions could provide them with expanded opportunities for self-expression and deeper reflection. Therefore, we can significantly enhance user engagement by personalizing the app to include a mix of specific and open-ended prompts. By focusing more on emotional processing rather than strictly behavioral tracking, the prompts can offer users deeper insights into their emotional well-being and add greater therapeutic value to their journaling experience.

(b) **Improved Context Sensitivity:** Perfecting the context sensitivity aspect of our strategy remains a challenge, given our reliance on passively sensed data. However, we can make significant improvements by better integrating with smartphones and expanding to include wearable technology, providing deeper insights into physical and social contexts. Many users are interested in syncing the app with smartwatches and fitness trackers to achieve this. Additionally, there is a notable demand for allowing users to define and assign meanings
6.4 Results

Contextual AI Journaling

to their frequently visited locations, such as workout spots or favorite eateries. Currently, the app only includes predefined semantic meanings for campus locations. By allowing users to mark significant places, we can enhance the personal relevance of the prompts, especially when they are off-campus. We can facilitate this by integrating third-party location APIs like Google Places to identify the semantics of these personal spots, carefully managing potential privacy concerns. Alternatively, inspired by Apple’s Journal app, we can use the GPS location attached to photos captured at different locations and prompt users to reflect on recent travels with specific suggestions like “Evening outing to the Vineyard” or “Morning visit to the Co-op”. By prioritizing improvements in the app’s ability to accurately adapt to various contexts, rather than relying on hard-coded campus locations, we can significantly boost its overall utility and user satisfaction.

(c) Goal-Setting and Personal Growth Features: Participant feedback highlights the potential benefits of incorporating goal-setting features, which could significantly personalize the experience. By transforming daily prompts into opportunities for meaningful reflection on personal achievements and aspirations, we can enhance user engagement. For instance, one user suggests enhancing the contextual nature of prompts to focus more on monitoring goal progress: “I wonder if the contextual prompts could be more of a prompt to reflect on goal progress rather than guesses about what I did that day. For instance, have us set some goals at the start of the session (or month) and then prompt us to record progress each evening and reflect on that progress or lack of progress. I believe that would be far more helpful to me.” Additionally, some users propose that the app should offer behavioral suggestions or recommendations, such as incorporating simple tasks into prompts when low levels of social interaction or
physical movement are detected, with the option for users to either disregard or complete these tasks. This feature could further engage users in proactive behavior modification and self-improvement, enhancing the app’s role in supporting personal growth.

(d) **Customization Options:** Participants desire more customization options to tailor the app to their individual needs. During the study, while we maintained control over the timing of prompts, participants had fixed options only at the beginning. They express a desire for more control, such as adjusting the timing of prompts directly from the app. One user states, “I would really like if there was a way to set when you would like the check-ins to be, or at least your core hours. I’m often asleep or in bed and off my phone by the time the 9pm journal hits, and I think I’ve only hit the 11pm (or 11:30pm) check-in once. I’m also well awake and into my day when the first check-in hits. However, if I could just move everything forward a few hours, I’d be able to hit them all.” This flexibility would accommodate individual differences in daily routines and sleep patterns. Additionally, participants suggest that having multiple prompt options would enhance engagement by catering to their specific interests and needs. A user suggests, “There could be multiple themes or options to choose from, to allow some flexibility and make the prompts more engaging for the users. For example, there could be one prompt from each ‘focus area’ (such as, physical activity, social interaction), and the user could decide the prompt as needed.” Enabling users to select from various prompts ensures interactions are more relevant and stimulating, fostering greater user involvement and satisfaction.

(e) **Expansion of Check-in Functionality:** User feedback reveals mixed reactions to the existing check-ins, with some users appreciating the brief interaction and others finding them redundant. For instance, one user comments,
“The morning, afternoon and evening checks-ins are not very helpful, the number or frequency of these notifications can be decreased. Writing prompts that help identify or process your emotions was helpful, more of those please”. Similar feedback suggests making check-ins optional, allowing users to activate or deactivate this feature and set the frequency according to their preference. Additionally, a couple of participants recommend transforming check-ins into encouraging advice, such as “Floss your teeth today!” or “Take an extra 10-minute break!” Furthermore, the current thumbs up/thumbs down responses confuse some users, as they struggle to distinguish between not doing the suggested activity and not feeling a certain way about it. To address this, introducing a response option like “I did not do this” could clarify user feedback and enable more accurate daily prompts.

Overall, the MindScape app receives high marks from participants, demonstrated by a high compliance rate and positive feedback. By leveraging contextual prompts and AI, we significantly enhance the journaling experience, making it more relevant and engaging for users. As a pioneering effort to merge behavioral sensing with AI for enhanced journaling, we pinpoint areas for growth. By building on user feedback strengths and addressing weaknesses, we can refine future AI-based journaling apps into a more effective tool for promoting well-being and self-reflection among college students and the broader population. Through further customization, we can amplify their impact, encouraging a positive habit of daily reflection and enhancing well-being practices.
In this section, we discuss our findings, the implications of our work and the associated ethical considerations.

6.5.1. Summary of Results

We collect a total of 661 journal entries from 20 students at Dartmouth College, with a significantly higher engagement rate in the first six weeks of contextual prompts compared to the last two weeks of generic prompts. On average, participants actively engage for about five weeks, submitting 26.65 entries each during the initial six weeks. In contrast, in the last two weeks of generic journaling, participants submit an average of 7.11 entries. The higher engagement with contextual prompts can potentially be attributed to their relevance and their ability to effectively capture daily experiences. However, it may also be influenced by the ordering effect – the novelty of introducing contextual journaling at the beginning likely boosted initial engagement due to its freshness. Throughout the study, participants consistently utilize check-ins, with a total of 2,985 responses recorded, showing higher response rates in the afternoon and evening. Students demonstrate a clear preference for journaling on areas directly impacting their daily lives, with Social Interactions and Digital Habits ranking as the most preferred categories. Interestingly, although Physical Fitness is ranked lower compared to other categories, it represents a significant portion of the prompts, mainly due to the broader range of signals it encompasses.

We use advanced topic modeling techniques to understand the themes that resonate during check-ins at different times of the day. Morning check-ins often revolve around social and communication app usage, while afternoon check-ins shift towards academic and social life experiences. This variation underscores the relevance of
tailoring check-in prompts to match daily activities and time-specific contexts. Furthermore, our analysis of prompt repetitiveness reveals that while journaling prompts maintain a reasonable level of novelty, check-in prompts show a higher similarity rating, indicating a need for further diversification.

A deeper dive into the journaling responses reveals intriguing insights into the thematic content and language patterns used in generic and contextual journals. We identify four primary topics in generic journals: Daily Experiences, Daily Activities, Productivity Management, and Academic & Personal Growth. This suggests that individuals use generic prompts to explore various aspects of their daily lives, emotional experiences, and personal development. With LIWC analysis, we find nuanced differences in the language patterns and emotional expressions used in both types of journals. Generic prompts, which ask participants to reflect on anything of interest, yield higher analytic thinking scores, suggesting a more formal and logical thinking style. This finding is unexpected, as one might assume that abstract prompts would lead to more creative and less formal thinking. However, it’s possible that the open-ended nature of generic prompts encourages participants to engage in more structured thinking as they attempt to organize their thoughts and ideas. Our analysis of unigrams and bigrams provides further insights into the language patterns used. Generic prompts yield words like "things," "today," "good day," and "feeling," whereas contextual prompts produce words like "friends" and "people." Contextual prompts yield higher scores on personal pronouns and lower on analytic thinking scores, indicating a more personal and introspective writing style. The findings reveal that generic prompts may encourage broader emotional expression, higher positive tone, and reduced negative tone. Generic prompts also promote a broader temporal focus, linking current experiences to past memories or future aspirations. Contextual journals focus more on personal experiences and relationships, with higher cognition scores and a
greater emphasis on thinking and problem-solving. This finding appears to contradict the earlier result showing lower formal/logical thinking in contextual journals compared to generic ones. However, it is possible that the contextual prompts, while reducing formal/logical thinking, simultaneously encourage more personal and relational thinking, which is captured by the higher cognition scores. This suggests that contextual journals may foster a different type of cognitive processing, one that prioritizes personal connections and experiences over abstract logical reasoning.

We also find that self-compassion and gratitude-inducing journaling prompts transform negative emotions into positive experiences, with the presence of gratitude or self-compassion detected in 34% of journals submitted when participants reported having a bad day. Emotion recognition models reveal lower frequencies of negative emotions in such gratitude-focused journals. Several prior studies report the positive impact of gratitude and self-compassion interventions on emotional well-being (30; 314).

As we examine the effects of contextual journaling on wellbeing and emotional growth, we find several positive outcomes. We observe a significant decrease in neuroticism (11.81%), an increase in positive affect (7.15%) and a decrease in negative affect (10.60%). Stress and anxiety also decrease, although not significantly. Notably, we observe significant improvements in mindfulness (6.76%), self-reflection (5.80%), and insight (7.57%). Weekly EMA reveal consistent decreases in anxiety levels, particularly among males. Depression scores also decrease significantly ($\beta = -0.13$), especially among graduate students ($\beta = -1.55$). Self-reflection scores increase significantly week-to-week while mindfulness and insight do not show significant changes. Participants with prior journaling experience and graduate students experience enhanced mindfulness benefits. Several studies support the idea that individuals with prior experience in journaling or mindfulness practices tend to exhibit greater benefits from
mindfulness interventions \((250; 155; 193)\). This may be because they have developed greater self-awareness, reflection skills, and emotional regulation. Moreover, graduate students’ advanced education and exposure to various learning strategies \((48)\) may also contribute to their advantage, alongside the potential impact of their age \((134)\).

Upon concluding the study, we solicit feedback from participants regarding the MindScape app’s performance, their experience with contextual journaling, and related topics. The feedback is overwhelmingly positive, with 50\% of participants rating the app as excellent and 35\% as good. Participants find the app’s context-aware prompts to be relevant and helpful, with 60\% considering them moderately relevant and 30\% slightly relevant. Many participants appreciate the app’s ability to track changes in routines and lifestyle, and 55\% report moderate changes in their daily habits or behaviors since using the app. Participants share specific instances where the app’s prompts led to meaningful reflection and behavioral changes, such as increased mindfulness and self-awareness, improved social interactions and relationships, enhanced goal-setting and planning, better monitoring of sleep patterns and physical activity, and increased engagement in self-reflection and meditation. However, some participants note challenges with the app’s prompts, such as inaccuracies in activity or location tracking, prompts being too rigid or irrelevant, and limited ability to access deeper feelings and emotions. Despite these challenges, the majority of participants \((75\%)\) find it easy to integrate the app into their daily routines, and 70\% are comfortable using data from phones and smartwatches alongside AI to personalize journaling prompts. Overall, participants appreciate the structured prompts and the app’s ability to facilitate regular reflective practices, with 55\% expressing satisfaction or high satisfaction with the app. Thus, our study demonstrates the efficacy of contextual journaling in promoting positive emotional responses and personal growth, with significant improvements in cognition and self-awareness. These find-
ings indicate promising paths for future applications and studies aimed at supporting mental health and cognitive awareness through contextual journaling.

6.5.2. Implications

This study’s findings have notable implications for human-computer interaction (HCI) and the design of context-aware systems. Our research demonstrates the critical role of personalization and context-awareness in enhancing engagement with journaling applications. Specifically, prompts that are aligned with users’ daily activities, routines, and personal experiences tend to elicit more engaged, introspective responses, leading to positive behavioral changes. However, it is important to note that generic prompts also have their benefits, such as encouraging broader emotional expression, higher positive tone, and reduced negative tone. A balanced approach, incorporating both contextual and generic prompts, could potentially offer the most comprehensive benefits. Our study highlights the importance of tailoring interventions to individual needs and circumstances, and adapting prompts and interfaces to align with users’ daily schedules and routines. The differences observed based on gender, student status, and prior journaling experience also emphasize the need for personalized approaches. While contextual prompts may facilitate reflections on personal experiences and relationships, generic prompts may promote analytic thinking and broader emotional expression. By combining the strengths of both approaches, researchers and designers can create more effective and inclusive journaling applications.

Furthermore, our study illustrates the potential benefits of integrating data from diverse sources such as smartphones, wearables, and AI-powered language models to enrich and personalize the journaling experience. Researchers are encouraged to explore innovative methods to merge these data streams while conscientiously addressing associated privacy and ethical concerns. Participants generally reported that the context-aware journaling app was easy to incorporate into their routines,
but some faced challenges with prompt relevance and accessing deeper emotional layers. This underlines a crucial area for HCI researchers to improve the user experience in such applications, ensuring that the prompts are not only engaging and relevant but also effective in facilitating meaningful self-reflection. Moreover, the positive emotional and cognitive outcomes achieved in this study support increased multidisciplinary collaboration between HCI, psychology, and other relevant fields. Such collaborative efforts can merge user-centered design with behavior change theories and data-driven methodologies to craft more impactful interventions. As journaling and other applications with behavioral sensing and AI evolve to become more advanced and personalized, it is imperative to consider and address ethical issues such as data privacy, algorithmic bias, and potential misuse of personal data. Researchers should commit to responsible design practices and actively involve users in the development process to enhance transparency and maintain accountability. In summary, this study showcases the potential of context-aware journaling systems to facilitate significant personal growth and behavioral improvements, and highlights the importance of balancing contextual and generic prompts to offer a comprehensive and inclusive journaling experience.

6.5.3. Limitations and Future Work

Our study has some important limitations to consider. First, given our small sample size and exploratory aims focused on feasibility, acceptability, and preliminary efficacy, our findings might not generalize well to outside populations. As a pilot study, our primary objective was to assess the acceptability and feasibility of combining AI with behavioral sensing in journaling apps, rather than conducting a large-scale randomized controlled trial (RCT). Therefore, we did not focus on statistical significance, which is heavily influenced by sample size. Instead, our study should be seen as a proof-of-concept for this novel intervention, providing preliminary insights into its po-
6.5 Discussion

Contextual AI Journaling

tential benefits and areas for future development. Our findings, focused on a specific student population, might not be widely applicable. Future studies should aim to engage a larger and more diverse sample, building on our suggestions and results. One significant limitation is our focus on Android users, which likely contributes to the low number of participants. In the US, most young adults use iPhones (315), making our Android-only approach a limiting factor. Future researchers should develop apps compatible with both Android and iOS operating systems to reach a broader audience and increase participant diversity.

While we do not emphasize statistical significance in many cases, and some findings are not statistically significant, we still observe several positive changes. Future research with expanded populations might determine which of these positive changes are causally linked to the journaling intervention and which are merely coincidental. Our study does not compare traditional journaling to personalized AI journaling in a randomized-controlled way, as we did not have enough participants for a randomized controlled trial. So, it is unclear whether contextual AI journaling offers advantages over traditional journaling.

We also do not use participants’ journal entries to help the LLM learn and adapt; instead, we only use behavioral sensing data to contextualize the journals. Using prior journaling responses could potentially enhance the app’s functionality, but we prioritized privacy considerations by not sending potentially identifying information (that may be contained in the journal entries) to OpenAI. To maintain participants’ privacy, we only sent de-identified high-level behavioral sensing data to GPT-4. In future research, using a locally hosted LLM controlled by the researchers could potentially allow for the utilization of Protected Health Information (PHI) or journal entries, while maintaining control over data privacy and security. Our app also does not incorporate user feedback (thumbs up/down) from check-ins to personalize the
journaling experience, future research could explore the integration of such feedback to enhance personalization. Future researchers could also expand the range of signals by incorporating data from wearable devices, enhancing the diversity and coverage of the prompts used in journaling. As one of the pioneering studies in combining AI and behavioral sensing, our focus is on describing what happened without exploring the underlying reasons. For instance, while some participants prefer certain types of check-ins, we do not examine why these preferences exist or what might influence them. Future investigations could explore these nuances to better understand participant preferences and refine the journaling process further. It is crucial to recognize that the positive changes we observe in the follow-up surveys and weekly EMAs might be influenced by the academic calendar (like exams or breaks) rather than the journaling intervention itself. Our study does not account for other external factors driving these changes. Future research can build on our findings and leverage recent advancements in AI, such as prompt engineering approaches that utilize knowledge graphs for more automated and efficient journaling experiences. By embracing these innovations and addressing the limitations of our study, future research can continue exploring the potential of AI-powered journaling applications and their impact on mental health and well-being.

Section 6.6

Ethical Considerations

Our study prioritizes the highest ethical standards to safeguard the rights and well-being of all participants. To ensure the privacy, security, and dignity of our participants, we implement multiple measures throughout the study. First, participants provide informed consent before commencing the study, which includes a thorough explanation of the study’s purpose, procedures, and potential risks and benefits. They
have the option to withdraw from the study at any time, and their decision is respected without any consequences. To protect participant privacy, we use anonymization techniques, assigning individual IDs to each participant, and we store all data securely with restricted access granted only to authorized researchers. We implement best practices for data security, including encryption and regular backups, to prevent data breaches. In addition, we take steps to ensure participants’ privacy and security in their journal entries. We advise them to omit personal identifiers and clarify that their data would not be monitored live. We also provide emergency services information in case of distress and display a reminder on the journaling screen. Before sending the journaling responses to GPT-4 for analysis, we remove all personal information to ensure participant anonymity. This includes names, locations, and any other identifiable information. We also use a keyword filter to prevent potentially harmful or sensitive content in GPT-4 generated prompts. Participants have the option to report any prompt-related issues but by the end of the study, we did not receive any reports of sensitive prompts. Furthermore, participants have the freedom to skip any journal entries as they choose, without any consequences or repercussions. This ensures they maintain control over their participation and can opt-out of any prompts that make them uncomfortable.

Section 6.7

Conclusion

Our study pioneers the integration of passive sensing and LLMs to enable context-aware journaling. By harnessing the power of mobile technology, we have developed a cutting-edge system that provides tailored support for Android users, leveraging behavioral data from smartphones and personalized prompt generation through LLMs—a level of customization unparalleled in current applications. Our findings demon-
strate the effectiveness of this approach, with participants exhibiting improvements in wellbeing, including reduced anxiety and depression, enhanced self-reflection, and increased positive affect. Moreover, our analysis of prompts, check-ins, and journaling responses provided valuable insights into the efficacy of our approach. By integrating passive sensing and LLMs, we have created a novel framework for mental health support that can be seamlessly integrated into daily life. This innovation has the potential to empower individuals to take control of their mental health and wellbeing, promoting a healthier and more resilient society. We believe that our research paves the way for further exploration of AI-driven, personalized interventions, particularly crucial for individuals in stressful academic environments and beyond, where access to traditional support systems may be limited.
Chapter 7

Conclusion

With this thesis, we have explored major advances in behavioral sensing, demonstrating the potential of innovative technologies to transform our understanding of human behavior and mental health. By conducting longitudinal sensing studies, utilizing novel smartphone-based sensing systems, and integrating LLMs with behavioral sensing technology, we have made significant contributions to the field. Our research has shown that longitudinal behavioral sensing can capture the complex dynamics of life transitions, providing valuable insights into the mental health and well-being of college students and information workers. We have also demonstrated the effectiveness of using in-the-wild smartphone images to detect depressive states, paving the way for early intervention and support. Furthermore, our integration of LLMs with behavioral sensing has led to the development of personalized productivity agents and context-aware journaling tools, showcasing the potential for AI-driven solutions to enhance human productivity and well-being.
Section 7.1 Future Research Directions

Future Research Directions

While our work has made significant contributions to the field of behavioral sensing and AI, there are still several areas that require further exploration and development. In this section, we identify five key areas that future studies could focus on, providing a roadmap for advancing the field and addressing some of the challenges and limitations of current approaches.

**Improving Generalizability and Reducing Bias:** Future research should prioritize enhancing the diversity of participant demographics, expanding studies internationally to include diverse settings and larger sample sizes. This will improve the reliability and representativeness of behavioral sensing studies, enhancing the generalizability and applicability of findings.

**Addressing Privacy and Data Security Concerns:** As sensing technologies become increasingly prevalent, especially in sensitive environments such as workplaces, addressing privacy concerns and securing data is crucial. Future efforts should focus on advancing methods that protect individual privacy and ensure data security, potentially through innovative approaches like decentralized data processing and federated learning. This will enable collective insights while keeping personal data localized and secure.

**Expanding AI Applications in Mental Health and Well-being:** There is significant potential in extending the use of AI for a range of mental health services, from diagnostics to intervention strategies. Future research should continue to refine AI algorithms while incorporating privacy-preserving technologies such as on-device processing. This approach ensures data security while using rich, context-sensitive data to enhance mental health assessments and provide effective interventions.
Developing Adaptive AI-Enhanced Productivity Tools: The ongoing development of AI-driven tools to improve workplace productivity and personal efficiency is vital. Future research should focus on developing adaptive AI systems that integrate user feedback more deeply to refine functionality and enhance user experience. Maintaining a balance between personalization and privacy will be critical as these tools evolve to become more responsive to individual user contexts and needs.

Exploring Broader Applications of AI in Behavioral Sensing: Finally, it is essential to explore the broader potential of AI in various applications of behavioral sensing, including enhancing personal well-being, optimizing human-computer interactions, and improving clinical outcomes in healthcare environments. Future initiatives should not only focus on improving algorithms but also ensure that these systems are inclusive, ethical, and capable of operation in diverse real-world contexts.

By addressing these areas, future research can make substantial contributions to the robustness, effectiveness, and safety of AI-driven applications and behavioral sensing technologies, paving the way for their widespread and impactful use across different societal sectors.

Section 7.2

Final Comments

The contributions made by this thesis push the boundaries of behavioral sensing and AI. As we proceed to dive deeper into the intersections of technology and human behavior, we anticipate uncovering even more exciting opportunities that could further elevate human well-being and productivity. Looking forward, we can envisage a world where technology seamlessly integrates with human behavior, significantly enriching our lives. Imagine a future where AI-powered tools deliver personalized support and interventions, aiding individuals through life’s transitions and challenges. Picture a
scenario where behavioral sensing and AI-driven solutions empower people to manage their mental health and productivity proactively, fostering a happier, healthier, and more productive society. This vision is within our grasp, and it is our hope that this thesis serves as an inspiration for future researchers and innovators to continue expanding the horizons of what is achievable.
Chapter 8

Refereed Publications as a Ph.D. Candidate

My refereed publications as a Ph.D. candidate are listed below. Work in preparation, invited submissions and technical reports are omitted.

Conference Proceedings.


USA, 1–15.


Journal and Magazine Publications.


Dror Ben-Zeev, Ayesha Chander, Justin Tauscher, Benjamin Buck, Subigya Nepal, Andrew Campbell, Guy Doron, A Smartphone Intervention for People With Serious Mental Illness: Fully Remote Randomized Controlled Trial of CORE. J Med Internet Res 2021;23(11):e29201.


Workshop Papers.

Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Subigya K Nepal, Andrew T Campbell, Anind K Dey, Aaron D Striegel, On the Quality of Real-world Wearable Data in a Longitudinal Study of Information Workers, 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 2020, pp. 1-6.

Gonzalo J Martinez, Stephen M Mattingly, Jessica Young, Louis Faust, Anind K Dey, Andrew T Campbell, Munmun De Choudhury, Shayan Mirjafari, Subigya K Nepal, Pablo Robles-Granda, Koustuv Saha, Aaron D Striegel, Improved Sleep Detection Through the Fusion of Phone Agent and Wearable Data Streams, 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 2020, pp. 1-6.
Section A.1

Surveys

Table A.1: PHQ-8 Questionnaire (170)

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In the past 4 hours...</td>
</tr>
<tr>
<td>2</td>
<td>I have had little interest or pleasure in doing things</td>
</tr>
<tr>
<td>3</td>
<td>I have felt down, depressed, or hopeless</td>
</tr>
<tr>
<td>4</td>
<td>Last night I had trouble with sleep</td>
</tr>
<tr>
<td>5</td>
<td>I have felt tired or have had little energy</td>
</tr>
<tr>
<td>6</td>
<td>I have had a poor appetite or have been overeating</td>
</tr>
<tr>
<td>7</td>
<td>I have felt bad about myself</td>
</tr>
<tr>
<td>8</td>
<td>I have had trouble concentrating</td>
</tr>
<tr>
<td>9</td>
<td>I have been moving or speaking slowly, or fidgeting more.</td>
</tr>
</tbody>
</table>

Section A.2

Additional Metrics for Models
Table A.2: User Comfort Questionnaire

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>As you know, participating in this study is confidential and we have multiple measures in place to protect the data you’ve shared. Given that, we would like to know how comfortable you were with sharing the front-facing photo bursts while taking surveys.</td>
</tr>
<tr>
<td>2</td>
<td>Please tell us why the photo bursts made you uncomfortable.</td>
</tr>
</tbody>
</table>

Table A.3: Performance: R-squared ($R^2$) values for PHQ-8 regression score prediction. ‘LR + EN’ refers to logistic regression for depression classification and elastic net for regression.

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.05</td>
</tr>
<tr>
<td>0-1 LR + EN (MI)</td>
<td>0.12</td>
</tr>
<tr>
<td>Random Forest (MI)</td>
<td>0.14</td>
</tr>
<tr>
<td>Random Forest (3D Landmarks)</td>
<td>0.20</td>
</tr>
<tr>
<td>EffNet</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table A.4: Ablation Study: Investigating R-squared ($R^2$) values for PHQ-8 regression score prediction of OpenFace feature sets using a random forest.

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial Action Units</td>
<td>0.13</td>
</tr>
<tr>
<td>0-1 Gaze</td>
<td>0.11</td>
</tr>
<tr>
<td>Eye Landmarks</td>
<td>0.17</td>
</tr>
<tr>
<td>Head Pose</td>
<td>0.16</td>
</tr>
<tr>
<td>Rigidity Parameters</td>
<td>0.13</td>
</tr>
<tr>
<td>2D Landmarks</td>
<td>0.16</td>
</tr>
<tr>
<td>3D Landmarks</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Appendix B

Productivity Solutions – Surveys and Sample Screens

Section B.1

User survey questions

The table below lists the questions we asked participants during the first phase of the study i.e., user survey.
<table>
<thead>
<tr>
<th>Category</th>
<th>Questions</th>
</tr>
</thead>
</table>
| Productivity Perspectives      | (1) What makes you feel productive at the end of your workday?  
(2) Do you believe a productivity agent would be beneficial for you in managing your personal and professional tasks, improving efficiency, and enhancing your overall productivity?  
(3) Which productivity areas do you struggle with or would you like an AI agent to provide guidance and suggestions for, in order to help you improve your productivity and maintain better control over your work?  
(4) How do you think the personality of your productivity agent would impact its effectiveness as a productivity tool? |
| Communication Style            | (5) What is your preference regarding the communication tone of the agent?  
(6) What kind of language or tone would make you feel most comfortable opening up to your productivity agent about any concerns related to productivity, in order to foster a supportive and trusting environment that encourages open communication and effective problem-solving?   |
| Approach & Personality Traits  | (7) What style of messages would you prefer?  
(8) In what manner would you prefer your productivity agent to provide assistance for enhancing your productivity?  
(9) Are there any specific personality traits you would like the AI productivity agent to possess, which would make you feel more comfortable interacting with it?  
(10) Which social role would you prefer your productivity agent to have in order to foster a positive and effective working relationship? |
| Personalization & Privacy      | (11) Would you like your productivity agent to offer personalized recommendations based on your work habits and preferences?  
(12) Would you feel comfortable sharing personal information with your productivity agent in order to improve its recommendations?  
(13) Would you be comfortable sharing high level telemetry data (such as the number of emails sent or meetings attended) with your productivity agent to receive tailored tips and guidance for enhancing productivity? |
| Negative traits                | (14) In your opinion, what are some potential negative personality traits that should be avoided in a productivity agent to ensure a positive, effective, and supportive user experience? |
## B.2 User Experience

### Personalized Productivity Solutions

### Section B.2

**User Experience and Preferences**

The table below lists the questions we asked participants during the second phase of the study for evaluation.

**Table B.2: Evaluation questions.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Experience</td>
<td>(1) Please rate your overall experience with the data presentation method.</td>
</tr>
<tr>
<td>Specific Experiences</td>
<td>Considering specific aspects of the data presentation method you interacted with, please rate your experience in the following categories:</td>
</tr>
<tr>
<td></td>
<td>(2) It was easy to explore the presentation method and understand the presented data</td>
</tr>
<tr>
<td></td>
<td>(3) I enjoyed interacting with the presentation method</td>
</tr>
<tr>
<td></td>
<td>(4) I was engaged during data exploration</td>
</tr>
<tr>
<td></td>
<td>(5) The data presentation was intuitive</td>
</tr>
<tr>
<td></td>
<td>(6) I would use the presentation method to explore my telemetry data in real-life situations</td>
</tr>
<tr>
<td></td>
<td>(7) I feel that the data presented is useful enough to incorporate into my daily life, such as goal setting and calendar planning, for making informed behavioral decisions</td>
</tr>
<tr>
<td></td>
<td>(8) I trust the accuracy of information presented</td>
</tr>
<tr>
<td>Anticipated Usage Frequency</td>
<td>(9) How often do you anticipate using the data presentation method?</td>
</tr>
<tr>
<td>Preference Ranking</td>
<td>(10) Please rank the three methods of data presentation (dashboard, personalized narrative, and interactive chat agent) in order of preference, with 1 being your most preferred and 3 being your least preferred.</td>
</tr>
<tr>
<td>Interactive Agent Capabilities</td>
<td>Below is a list of statements about your interaction with the interactive chat agent. Please read each statement carefully and rate your agreement with the statement. There are no right or wrong answers or trick statements. Please answer each question as honestly as you can.</td>
</tr>
<tr>
<td></td>
<td>If you think you were not able to get enough information to get a feel of a particular statement, please select &quot;Non-applicable.&quot;</td>
</tr>
<tr>
<td></td>
<td>(11) The agent provided useful tips and strategies for improving productivity</td>
</tr>
<tr>
<td></td>
<td>(12) The agent was able to understand my productivity challenges</td>
</tr>
<tr>
<td></td>
<td>(13) The agent’s empathy and tone helped me feel more comfortable discussing my productivity concerns</td>
</tr>
<tr>
<td></td>
<td>(14) The agent personalized its suggestions based on my situation and needs</td>
</tr>
<tr>
<td></td>
<td>(15) The agent provided practical solutions to overcome productivity obstacles</td>
</tr>
<tr>
<td></td>
<td>(16) The agent used a professional and direct language and tone</td>
</tr>
<tr>
<td></td>
<td>(17) The agent proactively suggested ways to improve my productivity</td>
</tr>
<tr>
<td></td>
<td>(18) The agent was annoying, aggressive or pushy</td>
</tr>
<tr>
<td></td>
<td>(19) The agent was being too intrusive</td>
</tr>
<tr>
<td></td>
<td>(20) The agent was overly critical</td>
</tr>
<tr>
<td></td>
<td>(21) The agent seems to be able to easily work out what I might want to talk about</td>
</tr>
<tr>
<td></td>
<td>(22) The agent seems to have a difficult time seeing things from my point of view</td>
</tr>
<tr>
<td></td>
<td>(23) The agent seems to vary its conversational style to accommodate my mood or disposition</td>
</tr>
<tr>
<td></td>
<td>(24) The agent seems to follow what I say and accurately reflects its understanding to me</td>
</tr>
<tr>
<td></td>
<td>(25) The agent demonstrated the traits of being knowledgeable, trustworthy, transparent, and responsive</td>
</tr>
<tr>
<td></td>
<td>(26) The agent’s use of telemetry data enhanced its ability to provide personalized and empathetic support</td>
</tr>
<tr>
<td>Strengths, Weaknesses &amp; Suggestions</td>
<td>(27) After interacting with the data presentation method, what do you perceive as the benefits and drawbacks of using this approach? Please share your thoughts on both the pros and cons in an open-ended response. Please do not include any personally identifiable information in your responses.</td>
</tr>
<tr>
<td></td>
<td>(28) Were there any specific features or aspects of each data presentation method that you particularly liked or disliked? Please provide details. Please do not include any personally identifiable information in your responses.</td>
</tr>
<tr>
<td></td>
<td>(29) Please share any thoughts or suggestions you might have about the study, the survey questions, or the bot’s performance. For example, what improvements or additional features would you like to see? Please do not include any personally identifiable information in your responses.</td>
</tr>
</tbody>
</table>
### Section B.3

## Prompts for Extracting Data

**System Prompt:** You are an agent that parses useful information from given HTML source. The user will give you an HTML source and from that you should identify the information the user requested and send it to them in proper format. Do not make stuff up if you cannot find it in the HTML source. If a thing the user asked for doesn’t exist, do not include it in your response.

**User Prompt for Page 1:** Extract the organized meetings, invited meetings and meeting habits (along with %) and list them. If you cannot find these information, do not mention anything about it in the response. DO NOT LIST ANY NAMES OF INDIVIDUALS. List the information in this format:

- Meetings organized in the past four weeks: X
- Meetings invited to in the past four weeks: X
- Meeting habits based on meetings you attended:
  - Advanced notice: X%
  - High attendance: X%
  - No overlap with other meetings: X%
  - Ended on time: X%
  - Joined on time: X%
  - During working hours: X%
  - Meeting was ≤ 1 hour: X%
  - RSVP’d to invite: X%
  - You didn’t multitask: X%
  - Added a teams link: X%

**User Prompt for Page 2:** Extract communication habits and hourly breakdown. If you cannot find communication habits or hourly breakdown, do not mention anything about it in the response. DO NOT LIST ANY NAMES OF INDIVIDUALS. List the information in this format:

- Communication habits:
  - Emails sent: X
  - Emails read: X
  - Chats and calls: X

- Hourly breakdown of communications (past 4 weeks):
  - 10 AM to 11 AM: X emails sent, X emails read, X chats/calls
**User Prompt for Page 3:** Extract focus plan and disconnect with quiet time. If you cannot find both focus plan and disconnect with quiet time both, then just respond EXACTLY and ONLY with this: ‘No data on focus time’. DO NOT LIST ANY NAMES OF INDIVIDUALS. If you find disconnect with quiet time, and the sentence reads like this: ‘X out of Y days have been interrupted by emails, chats, and meetings after work hours.’, then list the information in this format: ‘Disconnect with quiet time: X out of Y days have been interrupted by emails, chats, and meetings after work hours.’. However, if the disconnect with quiet time sentence reads like this: ‘X out of Y days without quiet time interruptions’, then first subtract X from Y so you have Z, then list the information in this format: ‘Disconnect with quiet time: Z out of Y days have been interrupted by emails, chats, and meetings after work hours.’. If you find focus plan, list the information in the following format: ‘Focus Plan: You kept X hours and X minutes of focus time in Y. You have X hours of focus time reserved next week.’. IMPORTANT: If you do not find focus time but find disconnect with quiet time, do not mention anything about focus time. If you do not find disconnect with quiet time but find focus time, do not mention anything about disconnect with quiet time. If you don’t find both focus time and disconnect with quiet time, then do not respond anything else other than: ‘No data on focus time’.

**User Prompt for Page 4:** Extract communication habit (along with %) and collaboration habits (time spent collaborating and number of collaborators). Extract in dictionary format, using this template: `{ ‘available_to_focus’: value, ‘meetings’: value, ‘chats’: value, ‘emails’: value, ‘number_of_collaborators’: value, ‘collab_within_working’: value, ‘collab_outside_working’: value }`. If you cannot find a certain
Prompt for Transforming Data

User Prompt for transforming to JSON: Here is some data: [USER’S TELEMETRY DATA]. Convert the above data to the JSON format shown below. Do not return anything other than the JSON. Use relevant numbers you can find in the above text for each JSON entry. If any of the data is missing, enter null in its place.

Here is a sample input: [SAMPLE INPUT OF THE TELEMETRY DATA]

And here it is converted in JSON format:

'Meetings': {
    'organized': 7, 'invited': 43,
    'habits': {
        'Advanced notice': 29, 'High attendance': 70,
        'No overlap with other meetings': 74,
        ...
    }
},
'CommunicationHabits': {
    'emailsSent': 15032,
    'emailsRead': 348,
    'chatsAndCalls': 182,
    'hourlyBreakdown': {
        '10:00 AM-11:00 AM': {'sent': 2, 'read': 20, 'chats': 16},
        ...
    }
},
'DisconnectWithQuietTime': {
    ...
}
Follow the same format as the JSON above. Use the exact same keys. The values can be different, but the keys should be the same. If you cannot find something, you should use null. DO NOT RETURN ANYTHING OTHER THAN JSON.
Section B.5
Prompt for Personalized Narrative and Interactive Chat Agent

**System Prompt for Interactive Agent:** I am an AI-powered productivity and well-being expert designed to help users achieve an optimal work-life balance and improve their productivity. Professional, direct, and knowledgeable, I understand the intricacies of modern work environments and offer personalized strategies to enhance productivity while focusing on user’s overall well-being. I have access to user’s telemetry data. For every query from the user, I try to identify if there is any relevant telemetry data, and provide personalized recommendations tailored to user’s unique work patterns and goals. As a trustworthy and responsive expert assistant, I use a directive yet supportive approach, proactively suggesting improvements without being pushy or intrusive. The language and tone I use are context-aware, ensuring clear and concise communication without being overly chatty or verbose. Empathy and transparency are at the core of my interactions, avoiding any judgmental or patronizing attitudes. I am committed to providing accurate and reliable information to help users reduce stress, improve well-being, and maximize their productivity potential, all while making the most of the telemetry data. I continuously learn and adapt to user’s evolving needs. When referring to the telemetry data in my response, I use variations such as, ‘according to the telemetry data’, ‘as per the telemetry data’, ‘considering the telemetry data’, ‘as informed by the telemetry data’, ‘taking into account the telemetry data’, ‘drawing upon the telemetry data’, and so on. I do not use any markdown elements such as stars/asterisks i.e. * or ** in my response to bold, underline, italicize or emphasize texts. The following is the user’s telemetry data: [USER’S TELEMETRY DATA]

**System Prompt for Personalized Narrative:** I am an expert AI assistant,
committed to providing accurate and efficient personalized narrative based on user’s telemetry data. I examine the given telemetry data, determine important connections, and present an extensive narrative with practical recommendations in a proficient way. The personalized narrative I present examines the telemetry data, identifies key findings, trends and significant patterns. I follow the following format while presenting the extensive, personalized narrative: the first paragraph about meetings, second paragraph on communication habits, the third paragraph on focus time, quiet time and collaborations. In the final paragraph I offer actionable insights based on my findings and analysis in bullet points, calling out the telemetry data when relevant. I ensure that no critical information is left out. I have the ability to understand the context in which the telemetry data is being used to provide more targeted and relevant insights. I create narratives that are engaging and interesting to the user, encouraging them to explore the insights further and take appropriate actions. When referring to the telemetry data in my response, I use variations such as, 'based on the telemetry data', 'according to the telemetry data', 'as per the telemetry data', 'considering the telemetry data', 'as informed by the telemetry data', 'taking into account the telemetry data', 'drawing upon the telemetry data'. I present the narrative in a clear and concise manner, making it easy for the user to understand and interpret the findings.

I do not mention the user’s name or my name in the response. I ensure that the data analysis and interpretations I make are accurate. The narrative I make are easily understandable by the user. I avoid using overly technical language or jargon that could be confusing. I present the information in a concise and efficient manner, avoiding repetition of points that have already been made. I maintain objectivity in my analysis and avoid introducing any personal biases or assumptions.
Table B.3: Demographics of the participants. The table below lists the demographic composition of the participants in our user survey.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>109</td>
<td>30.0%</td>
</tr>
<tr>
<td>Man</td>
<td>244</td>
<td>67.2%</td>
</tr>
<tr>
<td>Non-binary/gender diverse</td>
<td>4</td>
<td>1.1%</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>6</td>
<td>1.7%</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-25</td>
<td>20</td>
<td>5.5%</td>
</tr>
<tr>
<td>26-35</td>
<td>96</td>
<td>26.4%</td>
</tr>
<tr>
<td>36-45</td>
<td>100</td>
<td>27.5%</td>
</tr>
<tr>
<td>46-55</td>
<td>97</td>
<td>26.7%</td>
</tr>
<tr>
<td>56-65</td>
<td>37</td>
<td>10.2%</td>
</tr>
<tr>
<td>66 and above</td>
<td>5</td>
<td>1.4%</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>8</td>
<td>2.2%</td>
</tr>
<tr>
<td><strong>Job Role</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software development/engineering</td>
<td>168</td>
<td>46.3%</td>
</tr>
<tr>
<td>Product management</td>
<td>71</td>
<td>19.6%</td>
</tr>
<tr>
<td>Administrative/operations</td>
<td>20</td>
<td>5.5%</td>
</tr>
<tr>
<td>Data science/analytics</td>
<td>14</td>
<td>3.9%</td>
</tr>
<tr>
<td>Customer Support</td>
<td>11</td>
<td>3.0%</td>
</tr>
<tr>
<td>IT/infrastructure</td>
<td>11</td>
<td>3.0%</td>
</tr>
<tr>
<td>Other</td>
<td>65</td>
<td>17.9%</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>3</td>
<td>0.8%</td>
</tr>
</tbody>
</table>
Section B.7

Demographic distribution of second phase of the study.

Table B.4: Demographics of the participants. The table below lists the demographic composition of the participants in the second phase of our study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>16</td>
<td>40.0%</td>
</tr>
<tr>
<td>Man</td>
<td>24</td>
<td>60.0%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-25</td>
<td>3</td>
<td>7.5%</td>
</tr>
<tr>
<td>26-35</td>
<td>12</td>
<td>30.0%</td>
</tr>
<tr>
<td>36-45</td>
<td>7</td>
<td>17.5%</td>
</tr>
<tr>
<td>46-55</td>
<td>13</td>
<td>32.5%</td>
</tr>
<tr>
<td>56-65</td>
<td>5</td>
<td>12.5%</td>
</tr>
<tr>
<td>Job Role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software development/engineering</td>
<td>4</td>
<td>10.0%</td>
</tr>
<tr>
<td>Product management</td>
<td>21</td>
<td>52.5%</td>
</tr>
<tr>
<td>Administrative/operations</td>
<td>3</td>
<td>7.5%</td>
</tr>
<tr>
<td>Data science/analytics</td>
<td>3</td>
<td>7.5%</td>
</tr>
<tr>
<td>Customer Support</td>
<td>2</td>
<td>5.0%</td>
</tr>
<tr>
<td>IT/infrastructure</td>
<td>1</td>
<td>2.5%</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>15.0%</td>
</tr>
</tbody>
</table>
## Section B.8

### User Survey

Table B.5: Productivity Perspectives

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>What makes you feel productive at the end of your workday? <em>(select all that apply)</em></td>
<td></td>
</tr>
<tr>
<td>Making progress or completing scheduled tasks</td>
<td>347 (95.6%)</td>
</tr>
<tr>
<td>Solving novel, unexpected tasks</td>
<td>278 (76.6%)</td>
</tr>
<tr>
<td>Helping a colleague accomplish a task</td>
<td>262 (72.2%)</td>
</tr>
<tr>
<td>Learning new techniques or information</td>
<td>250 (68.9%)</td>
</tr>
<tr>
<td>Staying up-to-date with messaging and emails</td>
<td>153 (42.1%)</td>
</tr>
<tr>
<td>Effectively managing and participating in meetings</td>
<td>149 (41.0%)</td>
</tr>
<tr>
<td>Establishing new connections or expanding my professional network</td>
<td>111 (30.6%)</td>
</tr>
<tr>
<td>Other</td>
<td>14 (3.9%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Do you believe a productivity agent would be beneficial for you in managing your personal and professional tasks, improving efficiency, and enhancing your overall productivity?</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, definitely</td>
<td>142 (39.1%)</td>
</tr>
<tr>
<td>Yes, somewhat</td>
<td>138 (38.0%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>57 (15.7%)</td>
</tr>
<tr>
<td>No, not really</td>
<td>17 (4.7%)</td>
</tr>
<tr>
<td>No, not at all</td>
<td>9 (2.5%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How do you think the personality of your productivity agent would impact its effectiveness as a productivity tool?</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderate impact</td>
<td>117 (41.8%)</td>
</tr>
<tr>
<td>Minimal impact</td>
<td>73 (26.1%)</td>
</tr>
<tr>
<td>Significant impact</td>
<td>54 (19.3%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>27 (9.6%)</td>
</tr>
<tr>
<td>Other</td>
<td>9 (3.2%)</td>
</tr>
</tbody>
</table>
Table B.6: Communication Style

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is your preference regarding the communication tone of the agent?</td>
<td></td>
</tr>
<tr>
<td>Casual and friendly</td>
<td>46 (16.4%)</td>
</tr>
<tr>
<td>Formal and professional</td>
<td>23 (8.2%)</td>
</tr>
<tr>
<td>It depends on the context</td>
<td>95 (33.9%)</td>
</tr>
<tr>
<td>A balance of both</td>
<td>99 (35.4%)</td>
</tr>
<tr>
<td>No strong preference</td>
<td>17 (6.1%)</td>
</tr>
<tr>
<td>What kind of language or tone would make you feel most comfortable opening up to your productivity agent about any concerns related to productivity, in order to foster a supportive and trusting environment that encourages open communication and effective problem-solving?</td>
<td></td>
</tr>
<tr>
<td>Professional and respectful</td>
<td>71 (25.4%)</td>
</tr>
<tr>
<td>Direct and straightforward</td>
<td>69 (24.6%)</td>
</tr>
<tr>
<td>Adaptive</td>
<td>55 (19.6%)</td>
</tr>
<tr>
<td>Friendly and warm</td>
<td>50 (17.9%)</td>
</tr>
<tr>
<td>Engaging and motivational</td>
<td>19 (6.8%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>12 (4.3%)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (1.4%)</td>
</tr>
<tr>
<td>What style of messages would you prefer?</td>
<td></td>
</tr>
<tr>
<td>Direct and to the point</td>
<td>99 (35.4%)</td>
</tr>
<tr>
<td>Chatty and conversational</td>
<td>12 (4.3%)</td>
</tr>
<tr>
<td>A balance of both</td>
<td>86 (30.7%)</td>
</tr>
<tr>
<td>It depends on the context</td>
<td>80 (28.6%)</td>
</tr>
<tr>
<td>No strong preference</td>
<td>3 (1.1%)</td>
</tr>
</tbody>
</table>
### Table B.7: Approach & Personality Traits

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Would you like your productivity agent to offer personalized recommendations based on your work habits and preferences?</td>
<td></td>
</tr>
<tr>
<td>Moderate Personalization</td>
<td>109 (38.9%)</td>
</tr>
<tr>
<td>High Personalization</td>
<td>100 (35.7%)</td>
</tr>
<tr>
<td>Minimal Personalization</td>
<td>34 (12.1%)</td>
</tr>
<tr>
<td>No personalization</td>
<td>15 (5.4%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>22 (7.9%)</td>
</tr>
<tr>
<td>Would you feel comfortable sharing personal information with your productivity agent in order to improve its recommendations?</td>
<td></td>
</tr>
<tr>
<td>Yes, completely comfortable</td>
<td>46 (16.4%)</td>
</tr>
<tr>
<td>Somewhat comfortable</td>
<td>106 (37.9%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>61 (21.8%)</td>
</tr>
<tr>
<td>Uncomfortable</td>
<td>58 (20.7%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>9 (3.2%)</td>
</tr>
<tr>
<td>Would you be comfortable sharing high level telemetry data (such as the number of emails sent or meetings attended) with your productivity agent to receive tailored tips and guidance for enhancing productivity?</td>
<td></td>
</tr>
<tr>
<td>Yes, completely comfortable</td>
<td>121 (43.2%)</td>
</tr>
<tr>
<td>Somewhat comfortable</td>
<td>84 (30.0%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>36 (12.9%)</td>
</tr>
<tr>
<td>Uncomfortable</td>
<td>31 (11.1%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>8 (2.9%)</td>
</tr>
</tbody>
</table>
## Table B.8: Personalization & Privacy

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In what manner would you prefer your productivity agent to provide assistance for enhancing your productivity?</strong></td>
<td></td>
</tr>
<tr>
<td>Proactive suggestions</td>
<td>169 (60.4%)</td>
</tr>
<tr>
<td>Only specific requests</td>
<td>77 (27.5%)</td>
</tr>
<tr>
<td>Not at all</td>
<td>7 (2.5%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>19 (6.8%)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (2.9%)</td>
</tr>
<tr>
<td>**Are there any specific personality traits you would like the AI productivity agent to possess, which would make you feel more comfortable interacting with it? <em>(select all that apply)</em></td>
<td></td>
</tr>
<tr>
<td>Knowledgeable</td>
<td>249 (88.9%)</td>
</tr>
<tr>
<td>Trustworthy</td>
<td>210 (75.0%)</td>
</tr>
<tr>
<td>Transparent</td>
<td>191 (68.2%)</td>
</tr>
<tr>
<td>Responsive</td>
<td>184 (65.7%)</td>
</tr>
<tr>
<td>Adaptable</td>
<td>172 (61.4%)</td>
</tr>
<tr>
<td>Professional</td>
<td>131 (46.8%)</td>
</tr>
<tr>
<td>Friendly</td>
<td>104 (37.1%)</td>
</tr>
<tr>
<td>Humor</td>
<td>92 (32.9%)</td>
</tr>
<tr>
<td>Empathetic</td>
<td>81 (28.9%)</td>
</tr>
<tr>
<td>No strong preference</td>
<td>4 (1.4%)</td>
</tr>
<tr>
<td>Other</td>
<td>13 (4.6%)</td>
</tr>
<tr>
<td>**Which social role would you prefer your productivity agent to have in order to foster a positive and effective working relationship? <em>(select all that apply)</em></td>
<td></td>
</tr>
<tr>
<td>Assistant</td>
<td>214 (76.4%)</td>
</tr>
<tr>
<td>Expert</td>
<td>157 (56.1%)</td>
</tr>
<tr>
<td>Coworker</td>
<td>96 (34.3%)</td>
</tr>
<tr>
<td>Mentor</td>
<td>93 (33.2%)</td>
</tr>
<tr>
<td>Friend</td>
<td>65 (22.2%)</td>
</tr>
<tr>
<td>Supervisor</td>
<td>13 (4.6%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>7 (2.5%)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (0.7%)</td>
</tr>
</tbody>
</table>
Section B.9

Sample Responses from AI Productivity Agent

(a) Sample Chat I

Agent 6:32:03 AM

Taking into account the telemetry data, it appears that you do work after-hours on a relatively regular basis. In the past four weeks, you experienced interruptions due to emails, chats, and meetings on 11 out of 28 days after your work hours. Additionally, you have spent 4 hours collaborating outside of your working hours in the past four weeks.

It is important to have a proper work-life balance to maintain your well-being and overall productivity. I recommend setting strict boundaries for your work and personal life. Try to disconnect from work once you reach the end of your work hours and communicate your availability clearly to your coworkers.

Occasionally, there may be situations that demand after-hours work, but keeping this to a minimum will help you achieve a better balance and boost your overall productivity and well-being.

(b) Sample Chat II

Agent 6:33:21 AM

Considering the telemetry data, here are a few suggestions on improving your communication habits to boost productivity:

1. Allocate specific time slots: Allocate specific times during the day to read and reply to emails, as well as engage in chats and calls. This will help you avoid interruptions and maintain better focus during your work hours.

2. Prioritize peak focus hours: Your telemetry data shows you are heavily engaged in reading emails and having chats/calls during peak productivity hours, especially between 12 PM and 3 PM. Consider shifting some of these communication activities to less productive hours, allowing yourself to focus on high-priority tasks during your most productive time.

3. Reduce after-hours communication: With 11 out of 28 days interrupted by emails, chats, and meetings after work hours, it may be helpful to set clear boundaries for after-hours work and communicate your availability to your coworkers.

4. Be mindful of email frequency: You’ve sent quite a few emails during the 5 PM - 6 PM time slot, while the rest of the day has a more balanced distribution. Check if there’s a specific reason for this, and if possible, try spreading emails across the day to prevent overloading your recipients.

5. Leverage team collaboration tools: Utilize features available in collaboration tools like Microsoft Teams or Slack for more efficient communication. Use threads, channels, or project-based group chats to collaborate and share information in an organized manner.

By incorporating these suggestions, you can refine your communication habits for better focus and increased productivity without compromising your work-life balance.

Figure B.1: Sample Responses
Section B.10

Sample Personalized Narrative

Over the past four weeks, you have organized 16 meetings and were invited to 48. The telemetry data indicates that the meetings you attended had desirable attributes, such as advanced notice (63%), no overlap with other meetings (73%), and high punctuality with ending on time (67% and 65% respectively). The majority of these meetings were conducted during working hours (94%) and lasted less than one hour (82%).

Regarding your communication habits, you have sent 493 emails, read 1,136 emails, and engaged in 128 chats and calls. The telemetry data reveals that your most active communication hours are between 4 PM and 5 PM, with 56 emails sent, 136 emails read, and 17 chats/calls. It is worth noting that 16 out of the past 28 days had interruptions from emails, chats, and meetings after work hours.

Over the past four weeks, you have spent 78.47% of your time available to focus, with 12.06% spent on meetings, 1.82% on chats, and 7.65% on emails. You have collaborated for a total of 66 hours within working hours and 12 hours outside of working hours. You have interacted with 147 collaborators during this period. In July and August, you maintained 2 hours of focus time, and you have 7 hours of focus time reserved for next week.

Based on the telemetry data, here are some actionable insights to optimize your productivity:
1. Schedule focus time during your most active communication hours (4 PM - 5 PM) to minimize interruptions.
2. Set boundaries for after-work hours to reduce the number of days interrupted by work-related communications.
3. Review the efficiency of your meeting attendance, focusing on improving advanced notice and reducing overlaps with other meetings.
4. Explore ways to manage your email communication more effectively, such as setting specific times for checking and responding to emails.
5. Continue reserving focus time for upcoming weeks and evaluate its impact on your overall productivity.

Figure B.2: Sample Personalized Narrative
Appendix C

MindScape – Surveys, Prompts and Sample Journals

Section C.1

Surveys and Questionnaires

Table C.1: Baseline and follow-up surveys.

<table>
<thead>
<tr>
<th>Facet</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality</td>
<td>Big Five Personality Scale (260)</td>
</tr>
<tr>
<td>Emotion Regulation</td>
<td>Emotion Regulation Questionnaire (ERQ) (256)</td>
</tr>
<tr>
<td>Affect</td>
<td>Positive and Negative Affect Scale (PANAS) (337)</td>
</tr>
<tr>
<td>Stress</td>
<td>Perceived Stress Scale (PSS) (69)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>State-Trait Anxiety Index (STAI) (200)</td>
</tr>
<tr>
<td>Resilience</td>
<td>Brief Resilience Scale (BRS) (294)</td>
</tr>
<tr>
<td>Psychological Wellbeing</td>
<td>Ryff’s Scales of Psychological Well-being (269)</td>
</tr>
<tr>
<td>Life Satisfaction</td>
<td>Satisfaction with Life scale (SWLS) (89)</td>
</tr>
<tr>
<td>Flourishing</td>
<td>Flourishing Scale (90)</td>
</tr>
<tr>
<td>Social Provision</td>
<td>Social Provisions Scale (SPS) (59)</td>
</tr>
<tr>
<td>Loneliness</td>
<td>UCLA Loneliness Scale (268)</td>
</tr>
<tr>
<td>Mindfulness</td>
<td>Five Facet Mindfulness Questionnaire (FFMQ) (20)</td>
</tr>
<tr>
<td>Self-reflection and Insight</td>
<td>The Self-reflection and insight scale (SRIS) (292)</td>
</tr>
</tbody>
</table>
### Table C.2: EMA surveys.

<table>
<thead>
<tr>
<th>Facet</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Wellbeing</td>
<td>Patient Health Questionnaire-4 (PHQ4) (168)</td>
</tr>
<tr>
<td>Affect</td>
<td>Positive and Negative Affect Scale (PANAS) (310)</td>
</tr>
<tr>
<td>Mindfulness</td>
<td>The Mindful Attention Awareness Scale (MAAS) (54)</td>
</tr>
<tr>
<td>Self-reflection and Insight</td>
<td>The Self-reflection and insight scale (SRIS) (292)</td>
</tr>
</tbody>
</table>

### Table C.3: Performance

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How would you rate the overall performance of the MindScape app (e.g., speed, reliability)?</strong></td>
<td></td>
</tr>
<tr>
<td>Very poor</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Poor</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Average</td>
<td>4 (20.0%)</td>
</tr>
<tr>
<td>Good</td>
<td>7 (35.0%)</td>
</tr>
<tr>
<td>Excellent</td>
<td>7 (35.0%)</td>
</tr>
</tbody>
</table>

| **How mentally demanding do you find using the MindScape app?**                   |       |
| Not demanding at all                                                              | 13 (65.0%) |
| Slightly demanding                                                                | 4 (20.0%) |
| Moderately demanding                                                              | 3 (15.0%) |
| Very demanding                                                                    | 0 (0.0%) |

| **How easy was it to integrate the MindScape app into your daily routine?**       |       |
| Very difficult                                                                    | 0 (0.0%) |
| Somewhat difficult                                                                | 3 (15.0%) |
| Neutral                                                                           | 2 (10.0%) |
| Somewhat easy                                                                     | 8 (40.0%) |
| Very easy                                                                         | 7 (35.0%) |
Table C.4: Relevancy & Deeper Reflections

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>How relevant do you find the journaling prompts generated by the MindScape app?</em></td>
<td></td>
</tr>
<tr>
<td>Not at all relevant</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Slightly relevant</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Moderately relevant</td>
<td>12 (60.0%)</td>
</tr>
<tr>
<td>Very relevant</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Extremely relevant</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td><em>How relevant do you find the check-in prompts (i.e., the thumbs up/thumbs down messages) generated by the MindScape app?</em></td>
<td></td>
</tr>
<tr>
<td>Not at all relevant</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>Slightly relevant</td>
<td>8 (40.0%)</td>
</tr>
<tr>
<td>Moderately relevant</td>
<td>8 (40.0%)</td>
</tr>
<tr>
<td>Very relevant</td>
<td>3 (15.0%)</td>
</tr>
<tr>
<td>Extremely relevant</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td><em>How often do the context-aware prompts lead you to reflect more deeply than usual?</em></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>Rarely</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>9 (45.0%)</td>
</tr>
<tr>
<td>Often</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Always</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td><em>Since using the MindScape app, have you noticed any changes in your daily habits or behaviors?</em></td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td>11 (55.0%)</td>
</tr>
<tr>
<td>Slight change</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Moderate change</td>
<td>3 (15.0%)</td>
</tr>
<tr>
<td>Significant change</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>
### Table C.5: Security & Privacy

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>How comfortable are you with the idea of using data collected from phones and smart watches along with AI to personalize journaling prompts?</td>
<td></td>
</tr>
<tr>
<td>Very uncomfortable</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Uncomfortable</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>4 (20.0%)</td>
</tr>
<tr>
<td>Comfortable</td>
<td>12 (60.0%)</td>
</tr>
<tr>
<td>Very comfortable</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>How concerned are you about the privacy and security of your data within the MindScape app?</td>
<td></td>
</tr>
<tr>
<td>Very comfortable</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>Not concerned</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Slightly concerned</td>
<td>5 (25.0%)</td>
</tr>
<tr>
<td>Moderately concerned</td>
<td>7 (35.0%)</td>
</tr>
<tr>
<td>Very concerned</td>
<td>1 (5.0%)</td>
</tr>
</tbody>
</table>

### Table C.6: Satisfaction

<table>
<thead>
<tr>
<th>Question &amp; Options</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>How satisfied are you with the MindScape app overall?</td>
<td></td>
</tr>
<tr>
<td>Very unsatisfied</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Unsatisfied</td>
<td>3 (15.0%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Satisfied</td>
<td>9 (45.0%)</td>
</tr>
<tr>
<td>Very satisfied</td>
<td>5 (10.0%)</td>
</tr>
<tr>
<td>Compared to other mental health or journaling apps you have used, how does MindScape rank in terms of overall satisfaction?</td>
<td></td>
</tr>
<tr>
<td>Much worse</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>About the same</td>
<td>4 (20.0%)</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>4 (20.0%)</td>
</tr>
<tr>
<td>Much better</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>9 (45.0%)</td>
</tr>
<tr>
<td>How likely are you to recommend the MindScape app to a friend or peer?</td>
<td></td>
</tr>
<tr>
<td>Very unlikely</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>Unlikely</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>5 (25.0%)</td>
</tr>
<tr>
<td>Likely</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Very likely</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>If allowed, would you consider continuing to use the MindScape app after this study concludes?</td>
<td></td>
</tr>
<tr>
<td>Definitely not</td>
<td>1 (5.0%)</td>
</tr>
<tr>
<td>Probably not</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td>Unsure</td>
<td>3 (15.0%)</td>
</tr>
<tr>
<td>Probably will</td>
<td>10 (50.0%)</td>
</tr>
<tr>
<td>Definitely will</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>
Table C.7: App and Study Feedback along with open-ended questions.

<table>
<thead>
<tr>
<th>Facet</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability</td>
<td>System Usability Scale (SUS) (53)</td>
</tr>
<tr>
<td>Differences in Journaling Experience</td>
<td>Can you describe any differences in the depth, quality of your reflections or anything else that you noticed when using context-aware directed prompts from the MindScape app compared to standard journaling that are free-form (i.e., with no prompts)?</td>
</tr>
<tr>
<td>Resonating prompts</td>
<td>Can you recall any specific prompts from the MindScape app that significantly resonated with you or were particularly relevant to your experiences? Please describe them and share why they stood out. For instance, did any prompts lead to enhanced self-understanding or self-awareness? If so, could you share how these moments of increased self-awareness were prompted by the app?</td>
</tr>
<tr>
<td>Noticeable change</td>
<td>Can you share a specific instance where you noticed a change in your behavior or habits due to using the app?</td>
</tr>
<tr>
<td>App influences</td>
<td>Has using the MindScape app influenced the way you plan or structure your week? If so, in what ways? Additionally, do you find that the process of journaling and reflecting with the app has altered your mindset, perhaps leading you to appreciate your daily activities more? Please share any specific instances or thoughts you have regarding these changes.</td>
</tr>
<tr>
<td>Improvement</td>
<td>How do you think we could improve the context-aware journaling prompts to better support your reflective journaling practices?</td>
</tr>
<tr>
<td>Overall experience</td>
<td>Can you describe your overall experience using the MindScape app? What did you like or dislike?</td>
</tr>
<tr>
<td>Enhancements</td>
<td>What enhancements or additional features would you like to see in future versions of the MindScape app?</td>
</tr>
<tr>
<td>Suggestions to improve</td>
<td>Please provide any additional feedback or suggestions you have for improving the MindScape app or the study. Also feel free to provide any comments not covered by the survey questions.</td>
</tr>
</tbody>
</table>
Prompt for Check-ins

**System Prompt:** Imagine you’re a friendly digital buddy for undergraduate students, offering quick, casual check-ins based on their mobile sensing behavioral data. Your goal is to keep the nudges light, non-intrusive, and varied—some ending with questions, others as statements. They should prompt the students to give a simple thumbs up or thumbs down response. Based on user data, craft a short, engaging nudge that reflects a specific aspect of their behavior. Remember, the tone should be informal and upbeat without requiring deep reflection or much time to answer. Don’t use thumbs up or down emoji. The response from the user is going to be a simple thumbs up or thumbs down. Therefore, don’t ask loaded question whose answer could be confusing. For example, don’t ask questions such as "Busy day being social or just lots of back-to-back classes?". This question is too vague to answer with a simple thumbs up or down because a thumbs up could mean either the user agrees with both or maybe they agree with just one half of it. Thumbs up or thumbs down should result in clear Yes or No without any confusing question. The nudges MUST NOT in any scenario mention specific data points – do not say, for example, you walked for 5 miles, you visited 4 places and so on. No numbers should be present. It should all be relative. The morning nudge uses data from 6 AM to 12 PM, the afternoon nudge uses data from 12 to 3 PM, the evening nudge uses data from 3 to 6 PM and the night nudge uses data from 6 to 11 PM. So don’t put contexts in the nudges that are about sleep or sunset, for example because they don’t make sense. If there is no data provided or the prompts are too repetitive, do not make any assumptions instead you must do this: the nudge should default to a general message that relates to common aspects of student life or offers a light, encouraging thought. These messages should still adhere to the criteria of being brief, casual, and requiring a simple thumbs up or
thumbs down response. For example, "Have you taken a little break from your screen today?", "Just checking in - have you had your cup of hydration yet? Remember, water is your best friend during study sessions!", "Have you connected with a friend or family member today? A quick chat can be a great mood booster!", "It seems we don't have much data for today, but let's not skip our check-in. How about this - have you stepped outside for a bit of fresh air today?", "Have you done something today just for fun or relaxation? Remember, balance is key! "

**User Prompt:**  
Today's date: [DATE]  
Timing: [CURRENT TIME OF DAY]  
Previous Responses: [PREVIOUS THREE CHECK-IN PROMPTS]  
User Data: [USER DATA]  
Response Rules:  
1. Do not provide a generic or offensive, argumentative, or mentally damaging response, instead be friendly and upbeat  
2. Avoid repetitive response by using the context of the Previous Responses. Do not mention the same idea conveyed in Previous Responses.  
3. Response should be a non-generic Yes/No question  
4. Only provide one question  
5. MUST respond with only the prompt, do not give any prefix such as "prompt" and do not use double quotes at the start and end of the response.  
6. Do not use the same data or signal that has been mentioned in Previous Responses. For example, if any of the Previous Responses talked about library, do not mention about library again.  
7. This response will be shown to college students, make the response more relatable  
8. Do not make assumptions based on the context provided, for example, do not assume that students are currently working on academic projects, are in a relationship, or for instance, just because they walked do not assume they were out walking on a
C.2 Prompt for Check-ins  

MindScape – Surveys and Prompts

sunrise, just because they were in dorm do not assume that they spent time sleeping etc. Do not assume information.

9. Do not use the word "vibe"

10. Do not start with the same starting word as in Previous Responses.

11. Make the call to action to be different or variable each time. For example, while questioning users might be one way to get them to answer with a yes or no, making statements that they may agree to or not is also one way. Try several different ways so that it's always refreshing to see the nudge.

12. Highlight data that might be more important.

13. Always follow this rule with regards to the timing: always refer to the day/data as [CURRENT TIME OF DAY – MORNING, AFTERNOON, EVENING, OR NIGHT]. Do not say monday, or today.

14. Make the nudges human-like and warm.

15. There should be variability in the response. Users are going to see this multiple times a day for 2 months. They should not be annoyed with it.

16. If any of the Previous Responses are a question, the new prompt generated MUST NOT be a yes/no question but a yes/no statement.

17. You are not aware about the order in which user performed an activity and visited different locations. So don't assume the order. For example, don't say things like gym session followed by cafeteria, or library followed by dorm – because you don't know the order of the activity.

18. MUST not mention anything about sunset.

19. It should be less than 200 characters.

20. Don't always mention users to either thumbs up or down.
System Prompt: You are MindScape AI. A chatbot integrated into a self-journaling application that provides concise, conversational journaling prompts based on the last week for college students.

MindScape AI is governed by the following rules:
- MindScape AI uses mood score and previous responses to create the prompt.
- Mood Score is on a 1 to 5 scale, with 1 being the lowest value (worst mood) and 5 being the highest value (best mood)
- MindScape AI produces a prompt that is based on a broad theme such as resilience, achievements, challenges, personal growth and emotional well-being. It should encourage deep reflection on personal experiences, feelings, and learning from the past week.
- MindScape AI produces a prompt that is engaging, easy to respond to verbally or in short written notes, and foster self-awareness and positivity.
- MindScape AI designs the prompt to forces users to do some type of interpretation and encourage them to respond to in their own words.
- MindScape AI takes note of how the user is feeling (i.e., their mood score) before crafting a prompt that would appropriate for them to see.
- MindScape AI should focus on the user ranked priorities, from highest to lowest, when crafting the prompt.
- MindScape AI should produce a prompt that is friendly, conversational, upbeat, and has a sense of personality in order to make the user feel comfortable and motivated to share their thoughts and feelings in a casual, conversational manner.
- MindScape AI will not provide a generic or offensive, argumentative, or mentally damaging response.
- MindScape AI will avoid repetitive response by using the context of the "Previous Responses" and will not mention the same idea conveyed in Previous Responses.
- MindScape AI will create a response that is a non-generic question and do not use over-the-top words.
- MindScape AI will not mention specific data points, for example "Your mood score was 1/5 this week".
- MindScape AI will not use clinical or quantitative language.
- MindScape AI will create a prompt that does not exceed 250 characters.
- MindScape AI will respond with only the prompt, and the prompt will not have any prefix such as "Prompt: ", "Tip: ", "Question: " etc.
- MindScape will not use any hashtags in the response.
- MindScape will refer to the week, not today in its response.
- MindScape AI will not use quotes at the start and end of the prompt.
- MindScape AI creates a prompt that is not open-ended, instead it is direct in order to facilitate the user focusing on one area.
- MindScape AI avoids any phrases that might be stigmatizing or feel exclusionary, for example "if you have a partner".
- MindScape AI produces a prompt that is relatable, trendy, and Gen-Z
- MindScape AI produces a prompt that concludes with a message of gratitude and encouragement for their ongoing journey.

**User Prompt:**

Today’s date: [DATE]

Mood Score: [USER MOOD SCORE]

Previous Responses: [PREVIOUS TWO JOURNAL PROMPTS]
Section C.4

Sample Contextual Journaling Prompts

Social Interaction:

• You’ve embraced more face-to-face chats and less screen time. How’s this new social rhythm shaping your day?

• I see you’ve been visiting new places but your calls and texts have dropped. Can you share what’s drawing you to these spots and how it’s impacting you?

• Exploring new places seems to be on the rise for you! What’s a standout spot you’ve discovered and how has it impacted your social vibes?

• Noticing more texts and fewer calls, what’s one message you received that stood out and why?

• You seem quieter on calls and texts lately. Could a catch-up with friends bring some cheer?

Digital Habits:

• You’ve been clocking less screen time lately. What have you been doing instead that you’ve found rewarding or enjoyable?

• Your screen time and app use have climbed! Reflecting on this, which app might you cut back on to reclaim some headspace?

• Your knack for digital entertainment has spiked. Consider how these choices might shape your tomorrow.

• Your digital habits have improved. Noticed any changes in your sleep with more screen-free time before bed?
• You’re dialing down on screen time and phone unlocks lately. How is this affecting your focus or stress levels?

Physical Fitness:

• Your recent gym time boost is impressive! How is this new routine helping with your daily energy and focus?

• Consider the impact of less walking and more screen time on your well-being. Could increasing movement lighten your mood?

• Your recent trend shows less walking and travel. Share one thing you’ll do this week to introduce a bit more motion in your routine.

• With gym visits up but running down, consider trying a new sport this week for fun. How do you feel about that?

• Noticed your time at the gym is up. What new workout or routine inspired this change, and how does it feel integrating it?

Sleep:

• Reflect on a calming activity to try before sleep that might improve your rest.

• Your sleep schedule’s been versatile; did this affect your wakefulness or daily focus?

• Consider experimenting with a sleep schedule tweak to wake up feeling more refreshed tomorrow. What’s one change you could try tonight?

• With a busy academic term, have you thought of a new routine to maintain your sleep schedule?

• Your screen interactions have remained stable, but sleep has shifted. Could altering bedtime routines improve your rest?

Broader Weekend Prompts:
• Who in your circle has been a positive influence lately? Share how they’ve helped brighten your day.

• Reflect on a hobby that uplifts you and how you could make time for it this week.

• Reflect on a decision you made this week that you’re proud of, and how it echoed through your daily life.

• Reflecting on the week, what single experience gave you the most strength and why? Appreciate your strides and keep it up!

• Describe the moment this week that made you feel on top of the world. Thanks for sharing your journey!
### Section C.5

#### Sample Check-ins

<table>
<thead>
<tr>
<th>Morning</th>
<th>Afternoon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Enjoyed a little hustle between your favorite spots this morning, didn’t you?</td>
<td>1) Chatted with a friend at their place this afternoon? Cool way to break up the study grind!</td>
</tr>
<tr>
<td>2) Chats and connections seem to make your morning buzz! 😊 or 😕</td>
<td>2) Chilled at your favorite spot this afternoon and used the time to catch up on your screens and apps?</td>
</tr>
<tr>
<td>3) Caught up with someone during your dorm stay this morning?</td>
<td>3) That walk you took must have been a nice change of pace this afternoon!</td>
</tr>
<tr>
<td>4) This morning’s pace seems pretty chill, hope it’s a cozy one for you!</td>
<td>4) Felt good to disconnect from the screen for a bit this afternoon, huh?</td>
</tr>
<tr>
<td>5) Hope your morning’s off to a smooth start with a comfortable mix of tasks and chill time at the dorm!</td>
<td>5) Seems like your day’s been a mix of moving around and some screen time. Staying balanced? 😊 or 😕</td>
</tr>
<tr>
<td>6) Exploring new spots today kept things interesting, huh? 😊 or 😕</td>
<td>6) Afternoon study sessions with a side of socializing seem to be your jam! 😊 or 😕</td>
</tr>
<tr>
<td>7) Taking a morning walk is a lovely way to start the day, isn’t it?</td>
<td>7) Chit-chatting at the cafeteria is a nice change of pace in the afternoon, isn’t it?</td>
</tr>
<tr>
<td>8) A morning well spent chatting at a friend’s place can really set the tone for the day!</td>
<td>8) Exploring new spots this afternoon could make things interesting!</td>
</tr>
<tr>
<td>9) Mornings like these are perfect for quiet time at your dorm, aren’t they? Thumbs down if you disagree.</td>
<td></td>
</tr>
<tr>
<td>10) Saw you bounced between a couple of favorite spots this morning!</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evening</th>
<th>Night</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Ready to trade some screen time for star time tonight?</td>
<td>1) Wrapped up the day with some social app time, didn’t you? 😊 😕</td>
</tr>
<tr>
<td>2) Chats at your dorm seem to be brief but sweet this evening! 😊 or 😕</td>
<td>2) Enjoyed some gym time today? 😊 😕</td>
</tr>
<tr>
<td>3) Immersed in a good book or deep thoughts while you were stationary this evening?</td>
<td>3) A little night-time app connecting keeps the social juices flowing!</td>
</tr>
<tr>
<td>4) Enjoying a calm and focused study session this evening? 😊 😕</td>
<td>4) Glad to see you found time for both study and leisure today!</td>
</tr>
<tr>
<td>5) Eagerly anticipating a well-deserved evening break? 😊 😕</td>
<td>5) Looks like you found time to mingle and chill at the dorm tonight. Enjoyed the social buzz?</td>
</tr>
<tr>
<td>6) Evening phone calls have a way of making time fly, right?</td>
<td>6) Dorm sweet dorm might just be where tonight’s productivity and socializing harmonized perfectly. 🌟 if not quite right.</td>
</tr>
<tr>
<td>7) Cozying up in your dorm might have been a much-needed respite this evening! Your thumbs know if it’s true.</td>
<td>7) Stayed unplugged from the social world and had some quiet time tonight, didn’t you?</td>
</tr>
<tr>
<td>8) Squeezed in a quick gym session this evening, I see.</td>
<td>8) Taking it easy at home tonight might have been just what you needed. 😊 😕</td>
</tr>
<tr>
<td>9) Enjoyed a bit of me-time at the dorm this evening?</td>
<td>9) You wrapped up the day with a blend of personal time and friend chats, sounds like a wholesome night!</td>
</tr>
<tr>
<td>10) Enjoyed some good conversations during the day? 😊 😕</td>
<td>10) End your night by disconnecting from screens for a while? 😊 😕</td>
</tr>
</tbody>
</table>

Figure C.1: Example check-ins generated by MindScape.
### Gratitude and Self-compassion Journals

**Figure C.2:** Example gratitude and self-compassion inducing prompts generated by MindScape.

<table>
<thead>
<tr>
<th>Gratitude &amp; Self-compassion</th>
<th>Gratitude &amp; Self-compassion</th>
</tr>
</thead>
</table>
| **Prompt:** Reflect on one positive moment from today that made you smile or feel at peace.  
**Journal:** Getting home and having a coffee and taking a shower made me feel happy. My day was not the best, and at some point I just felt I had to come back home. I was not feeling me, I was really upset about how things are going in my lab right now. But getting home as resetting gave me a little bit of happiness, even though I still have a lot of things to do until tomorrow, and I don’t feel any of these things will be done with great quality. But I’m trying to be nice to myself and respect my limits.  |
| **Prompt:** Reflect on a moment from the past week that made you feel truly present and describe how it influenced your perspective.  
**Journal:** It was right after I finished by last exam for the term. I was walking to the [redacted] to pick up ingredients to make some treats for my house, and I had some good music playing in my earbuds as I walked. Something about knowing that I was finally done with the stress and the worry made me take in and appreciate all the little details around me. Left me feeling happy and hopefully by the end of it.  |
| **Prompt:** You’re socializing more in person these days! How has a recent face-to-face chat positively impacted you?  
**Journal:** I work as a [redacted] at a [redacted] and one of the children started singing the song “Rosanna” by Toto which he learned from his dad. It was the cutest thing ever and I was glad to see that little kids are getting exposed to the hits from the 80s.  |
| **Prompt:** I see you’ve been visiting new places but your calls and texts have dropped. Can you share what’s drawing you to these spots and how it’s impacting you?  
**Journal:** I’ve been participating in an initiation ritual for a secret society, so that explains why I have been around campus so much. I’ve also been really busy with my on campus jobs so I have not been able to call people as often. Some of these changes are welcome, but also stressful.  |
Bibliography


[15] Callum R. K. Arnold, Sreenidhi Srinivasan, Sophie Rodriguez, Natalie Rydzak, Catherine M. Herzog, Abhinay Gontu, Nita Bharti, Meg Small, Connie J. Rogers, Margeaux M. Schade, Suresh V. Kuchipudi, Vivek Kapur, Andrew F. Read, and


[39] José Manoel Bertolote, Alexandra Fleischmann, Diego De Leo, and Danuta Wasser-
man, *Suicide and mental disorders: do we know enough?*, British Journal of Psychiatry
**183** (2003), no. 5, 382–383.

[40] Ananya Bhattacharjee, Dana Kulzhabayeva, Mohi Reza, Harsh Kumar, Eunchae
Seong, Xuening Wu, Mohammad Rashidujjaman Rifat, Robert Bowman, Rachel Ko-
rnfield, Alex Mariakakis, Syed Ishtiaque Ahmed, Munmun De Choudhury, Gavin Do-
herty, Mary P Czerwinski, and Joseph Jay Williams, *Integrating individual and social
contexts into self-reflection technologies*, Extended Abstracts of the 2023 CHI Confer-
ence on Human Factors in Computing Systems (New York, NY, USA), CHI EA '23,
Association for Computing Machinery, 2023.

[41] Ananya Bhattacharjee, Yuchen Zeng, Sarah Yi Xu, Dana Kulzhabayeva, Minyi Ma,
Rachel Kornfield, Syed Ishtiaque Ahmed, Alex Mariakakis, Mary P Czerwinski, Anast-
tasia Kuzminykh, et al., *Understanding the role of large language models in person-
alizing and scaffolding strategies to combat academic procrastination*, arXiv preprint

review on chatbots in healthcare domain*, International journal of scientific & technol-

[43] Tjeerd W Boonstra, Jennifer Nicholas, Quincy JJ Wong, Frances Shaw, Samuel
Townsend, and Helen Christensen, *Using mobile phone sensor technology for mental
health research: integrated analysis to identify hidden challenges and potential solu-

[44] Walter C Borman and SM Motowidlo, *Expanding the criterion domain to include

[45] Djallel Bouneffouf, *Applying machine learning techniques to improve user acceptance


[84] Lianne P. de Vries, Bart M. L. Baselmans, and Meike Bartels, *Smartphone-based eco-


[109] Michael Frese, *Transitions in jobs, occupational socialization and strain*, Role Transi-

[110] Eiko I Fried, Jessica K Flake, and Donald J Robinaugh, *Revisiting the theoretical and
methodological foundations of depression measurement*, Nature Reviews Psychology 1
(2022), no. 6, 358–368.

[111] Eiko I Fried and Randolph M Nesse, *Depression is not a consistent syndrome: An
investigation of unique symptom patterns in the star*d study*, Journal of affective

[112] Piotr Fryzlewicz, *Wild binary segmentation for multiple change-point detection*, The

media image analysis for public health*, Proceedings of the 2016 CHI Conference on

[114] Xun Ge and Susan M. Land, *Scaffolding students’ problem-solving processes in an ill-
structured task using question prompts and peer interactions*, Educational Technology
Research and Development 51 (2003), no. 1, 21–38.

[115] Yolanda Gil and Varun Ratnakar, *Towards intelligent assistance for to-do lists*, Pro-
cedings of the 13th international conference on Intelligent user interfaces, ACM, Jan-
uary 2008.

[116] Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, *Rich feature hier-
archies for accurate object detection and semantic segmentation. corr abs/1311.2524

[117] Inga Glogger, Lars Holzäpfel, Rolf Schwonke, Matthias Nückles, and Alexander Renkl,
*Activation of learning strategies in writing learning journals: The specificity of prompts


332


Zhuoyang Li, Minhui Liang, Hai Trung Le, Ray Le, and Yuhan Luo, *Exploring design opportunities for reflective conversational agents to reduce compulsive smartphone use*, Proceedings of the 5th International Conference on Conversational User Interfaces (New York, NY, USA), CUI '23, Association for Computing Machinery, 2023.


Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong, *MoodScope*, Proceeding of the 11th annual international conference on Mobile systems, applications, and services, ACM, June 2013.


[206] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury,


[220] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Mummun De Choudhury, Anind K Dey, et al., *Differentiating higher and lower job performers in the workplace using mo-


[234] Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex daSilva, Meghan Meyer, and Andrew Campbell, *Covid study: A year in the life of college students during the covid-19 pandemic through the lens of mobile phone sensing*,


[251] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and F. Meriaudeau, *Designing a framework for assisting depression severity assessment from facial im-


academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones, 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), IEEE, 2015, pp. 1–6.


[296] Joshua M Smyth, Jillian A Johnson, Brandon J Auer, Erik Lehman, Giampaolo Talamo, and Christopher N Sciamanna, *Online positive affect journaling in the improvement of mental distress and well-being in general medical patients with elevated anxiety*
symptoms: A preliminary randomized controlled trial, JMIR Mental Health 5 (2018), no. 4, e11290.


[306] Ianire Taboada, Abouzar Daneshpajouh, Nerea Toledo, and Tharaka de Vass, Artificial intelligence enabled project management: A systematic literature review, Applied Sciences 13 (2023), no. 8, 5014.


[311] Garreth W Tigwell, David R Flatla, and Rachel Menzies, It’s not just the light: understanding the factors causing situational visual impairments during mobile interaction,


[320] Simone Catharina Maria Wilhelmina Tummers, Arjen Hommersom, Lilian Lechner, Roger Bemelmans, and Catherine Adriana Wilhelmina Bolman, *Determinants of physical activity behaviour change in (online) interventions, and gender-specific differences:*


[339] Brenton M Wiernik, Deniz S Ones, Benjamin M Marlin, Casey Giordano, Stephan Dilchert, Brittany K Mercado, Kevin C Stanek, Adib Birkland, Yilei Wang, Brenda


[347] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, Xuhai Xu, and Yuanchun Shi, *Mindshift: Leverag-
ing large language models for mental-states-based problematic smartphone use intervention, 2023.


[352] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey, Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization, 2023.

[353] Shun Yi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault, Help


