MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention

Ruolan Wu
Tsinghua University
Beijing, Beijing, China
wurl21@mails.tsinghua.edu.cn

Chun Yu*
Tsinghua University
Beijing, Beijing, China
chunyu@tsinghua.edu.cn

Xiaole Pan
Tsinghua University
Beijing, Beijing, China
pxl22@mails.tsinghua.edu.cn

Yujia Liu
Tsinghua University
Beijing, Beijing, China
l-yj22@mails.tsinghua.edu.cn

Ningning Zhang
Tsinghua University
Beijing, Beijing, China
znn18@tsinghua.org.cn

Yue Fu
University of Washington
Seattle, Washington, USA
chrisfu@uw.edu

Yuhan Wang
Beijing University of Posts and Telecommunications
Beijing, Beijing, China
2020211730@bupt.cn

Zhi Zheng
Tsinghua University
Beijing, Beijing, China
georgezhengzhi@gmail.com

Li Chen
Tsinghua University
Beijing, Beijing, China
chenli19@mails.tsinghua.edu.cn

Qiaolei Jiang
Tsinghua University
Beijing, Beijing, China
qiaoleijiang@tsinghua.edu.cn

Xuhai Xu
Massachusetts Institute of Technology
Cambridge, MA, USA
xoxu@mit.edu

Yuanchun Shi
Tsinghua University
Beijing, Beijing, China
shiyc@tsinghua.edu.cn

ABSTRACT

Problematic smartphone use negatively affects physical and mental health. Despite the wide range of prior research, existing persuasive techniques are not flexible enough to provide dynamic persuasion content based on users' physical contexts and mental states. We first conducted a Wizard-of-Oz study (N=12) and an interview study (N=10) to summarize the mental states behind problematic smartphone use: boredom, stress, and inertia. This informs our design of four persuasion strategies: understanding, comforting, evoking, and scaffolding habits. We leveraged large language models (LLMs) to enable the automatic and dynamic generation of effective persuasion content. We developed MindShift, a novel LLM-powered problematic smartphone use intervention technique. MindShift takes users' in-the-moment app usage behaviors, physical contexts, mental states, goals & habits as input, and generates personalized and dynamic persuasive content with appropriate persuasion strategies. We conducted a 5-week field experiment (N=25) to compare MindShift with its simplified version (remove mental states) and baseline techniques (fixed reminder). The results show that MindShift improves intervention acceptance rates by 4.7-22.5% and reduces smartphone usage duration by 7.4-9.8%. Moreover, users have a significant drop in smartphone addiction scale scores and a rise in self-efficacy scale scores. Our study sheds light on the potential of leveraging LLMs for context-aware persuasion in other behavior change domains.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Problematic smartphone use, persuasion, large language model, mental model

ACM Reference Format:
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. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642790

1 INTRODUCTION

In recent years, the ubiquitous presence of smartphones has increased people's reliance on digital devices, resulting in problematic smartphone usage behaviors, *i.e.*, excessive or mindless usage with negative consequences [45, 76], especially among adolescents and young adults [67]. Prior studies suggest that problematic smartphone usage can detrimentally affect people in various areas such as efficiency (leading to diminished academic or work

*Corresponding author.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642790

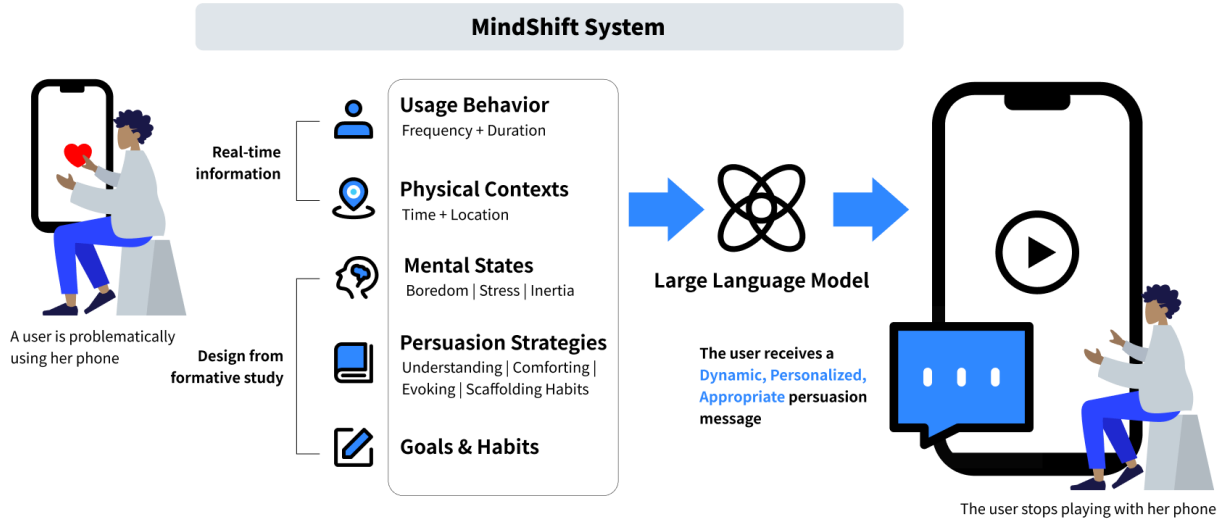


Figure 1: Overview of MindShift. When users exhibit problematic phone usage, MindShift actively collects data on phone usage behavior, physical contexts, and mental states, and uses the customized persuasion strategies we designed along with users' goals and habits, to generate prompts. Then, the large language model will generate a persuasion message. Finally, the persuasion will show up in users' phones to encourage more mindful usage.

performance [9, 26]), physical well-being (resulting in decreased sleep and activity levels [22, 66, 124, 125]), and mental health (manifesting as anxiety and depression [37, 39, 127]). Many individuals have recognized their problematic smartphone usage and sought to reduce their over-reliance on smartphones [42, 58].

A plethora of academic research and commercial products provide just-in-time (JIT) smartphone use interventions, intervening precisely when problematic use occurs [90]. These interventions fall into four categories based on their enforcement levels: (1) Self-monitoring: offering insights on phone usage patterns via notification or visualization, enhancing user awareness about their smartphone habits [5, 23, 57, 123]; (2) Reminders: countering immediate phone indulgence and promoting self-reflection through pop-up notifications [42, 98]; (3) Interaction friction: raising the effort needed to use the phone, thereby reducing its allure by introducing tasks like typing [93, 99, 128]; (4) Lockout: disabling the user's phone access for a specified duration [54, 55, 58, 72]. However, there are a few gaps among existing intervention techniques.

First, existing methods are limited to strike a balance between effective intervention engagement and good usability [85]. The first two categories rely on the user's self-control and are easily ignored, leading to low engagement and limited effectiveness [42, 57, 123]. The other two types are more restrictive, often causing user frustration due to reduced usability across different contexts [54]. Our approach utilizes reminder-based interventions with persuasive content to encourage reduced smartphone usage. Persuasion, typically through natural language to influence people's thoughts and behavior [20, 108, 114], is more effective with diverse and context-specific content [50, 51], as supported by recent studies highlighting the success of personalized [18, 36, 50, 60], context-aware [53, 58, 107, 116] interventions. However, most current reminders use repetitive, template-based content, reducing efficacy [8, 46, 128]. To overcome this, we employ Large Language Models (LLMs) [19, 94] to generate varied persuasive content. LLMs' reasoning ability provides a promising solution to infer users' current activities based on contextual information collected from smartphone sensors, such as time and location [16, 28, 56], enabling the creation of more relevant and effective intervention language.

Second, we identified the opportunity to leverage mental states associated with problematic smartphone use, an essential aspect of user contexts. While some current interventions use context-based strategies, like triggering interventions at specific times and locations [53, 58], they tend to focus mainly on external physical contexts, neglecting internal mental factors. Mental factors like stress and negative emotions are increasingly recognized as one key factor leading to problematic smartphone use [27, 78, 120, 121]. However, existing studies primarily address prolonged mental states rather than momentary contexts. Our study aims to bridge this gap by integrating an understanding of in-the-moment mental states into the intervention framework. We believe that a more holistic approach, considering both the physical and mental contexts, could enhance intervention effectiveness.

To address these gaps, we first conducted a Wizard-of-Oz study (N=12), followed by an interview study (N=10) to better understand users' mental states during problematic phone usage. Focusing on habitual usage (*i.e.*, ritualistic behavior, without a clear goal, such as passive social media content consumption) [43, 109], we summarized three major mental states to address: *boredom*, *stress*, and *inertia*. Building on the Dual Systems Theory [44] and the ERG (Existence, Relatedness, and Growth) Theory [15], we proposed four persuasion strategies: 1) understanding, 2) comforting, 3) evoking, and 4) scaffolding habits.

Integrating our persuasion strategies with LLMs, we designed and implemented MindShift (Figure 1), a new JIT intervention technique that can provide dynamic, personalized persuasion content based on user contexts. MindShift leverages LLMs' strong capability in commonsense comprehension and natural language generation [19, 83] to generate proper and effective persuasive content based on real-time information (phone usage behavior, physical contexts, and mental states) and long-term user states (user goals and habits), guided by persuasion strategies we designed.

To evaluate MindShift's effectiveness, we conducted a 5-week field experiment deploying our intervention to 25 participants. We compared MindShift against the baseline, a basic notification-based intervention that asked users to reflect on and report the purpose of their smartphone usage. Moreover, to assess the effect of the mental states factor, we compared MindShift against a simplified version, MindShift-Simple, that excludes the mental states factor from the LLM-based content generation.

Our study results indicate that MindShift and MindShift-Simple outperformed the baseline method on the intervention acceptance rate by 22.5% and 17.8%, respectively, with statistical significance. They also significantly reduce overall app opening frequency by 12.1% and 14.4% and app usage duration by 9.8% and 2.4%. Comparing MindShift and MindShift-Simple, including the mental state factor enhances the persuasion acceptance rate by 8.1% with statistical significance. Moreover, the subjective report data shows that participants using MindShift and MindShift-Simple experience a significant reduction in smartphone addiction scale (SAS) score (34.7% and 25.8% respectively) and an increase in the self-efficacy scale score (10.7% and 10.4% respectively).

Our paper makes the following contributions:

(1) We conducted a Wizard-of-Oz study and an interview study, uncovering three major mental states (*boredom*, *stress*, and *inertia)* during habitual smartphone use, which led us to design four persuasion strategies grounded in the Dual Systems Theory and the ERG Theory: *Understanding, Comforting, Evoking,* and *Scaffolding Habits.*
(2) We created MindShift, a novel persuasive intervention technique leveraging LLMs to generate dynamic and personalized persuasion content based on users' phone usage behavior, physical contexts, mental states, goals and habits, and appropriate persuasion strategies.
(3) We conducted a field experiment by deploying MindShift, demonstrating significant improvements in intervention acceptance rates and reduced smartphone use by Mindshift. Users' subjective feedback also corroborated these observations, validating the effectiveness of MindShift.

2 RELATED WORK

In this section, we define problematic smartphone use and habitual smartphone use (Sec. 2.1), explore the reasons behind engagement in problematic smartphone use (Sec. 2.2). We then briefly overview existing intervention techniques (Sec. 2.3). Finally, close to our work, we introduce behavior change persuasion techniques and their relationship with the emergence of LLMs (Sec. 2.4).

2.1 Problematic Smartphone Use

Many studies have explored the definition of problematic smartphone use, which can be broadly classified into two categories. The first category defines whether users exhibit addictive behaviors toward their phones. Some studies assess addictive behavior by measuring the level of user dependence on smartphones through questionnaires, such as the Smartphone Addiction Scale (SAS) and the Smartphone Addiction Inventory (SPAI) [64, 70]. The second category defines whether a specific instance of phone use is problematic. Growing research suggests that problematic smartphone use is determined not only by excessive use but also by the purpose and content of use in specific situations [39, 74, 75, 76, 102, 104]. Studies have indicated that phone use purposes can be categorized into (1) habitual use that is performed unconsciously and ritualistically usually without a specific goal [101], and (2) instrumental use with a specific task or goal in mind [76]. Existing research suggests that habitual use should be the primary target for intervention [76, 89, 101]. In this paper, we use SAS to measure users' level of addictive smartphone usage. We also distinguish users' phone use purpose and focus on intervening habitual use.

2.2 Understanding Problematic Smartphone Use through A Dual Systems Perspective

Understanding what leads to problematic smartphone use is essential for the design of effective persuasion strategies. The Dual Systems Theory [44, 49] has been used to explain the phone usage patterns [101]. This theory divides human cognitive activities into two types: System 1 (fast, intuitive, unconscious) and System 2 (slow, analytical, conscious). Problematic smartphone use is typically driven by System 1 [34], as it mainly involves unconscious, rapid responses and is easy to be guided by instant gratification. Research suggests that two key factors contribute to the failure of users to act on their goals: (1) limited ability of System 2 control; and (2) fluctuations of System 2 caused by emotional states and fatigue [78]. For the first factor, a growing amount of research suggests that the limited ability of control is attributed more to apps' deliberate design than to users themselves [11, 33, 75, 87, 106]. Recent work identifies types of attention-capture deceptive designs in digital interfaces, such as neverending autoplay and infinite scroll [87]. For the second factor, some findings suggest that mental states play a significant role in habitual smartphone use [101, 120, 121] and previous research has identified external contexts, such as social awkwardness [104, 118], that may trigger the habitual use. However, there is limited research exploring what specific kinds of users' in-the-moment mental states behind habitual use. In our work, we pinpoint the major mental states linked to habitual use and propose corresponding persuasion strategies.

2.3 Problematic Smartphone Use Intervention

Existing problematic smartphone use intervention techniques fall into two groups: external interventions that monitor and limit use, and internal interventions that change the interface itself.

External intervention can be roughly divided into four categories based on enforcement level: The first category provides users with information about their behavior such as visualization of usage [8, 23, 42, 57, 73, 77, 92, 97, 123], requiring users to view it themselves

to increase awareness of phone usage. The second actively sends reminders to users to provoke their reflection such as reminding users of their daily goals [42, 79], informing their usage time [8] or the number of opens [105]. This category presents text to users and, therefore, serves as a persuasion. The third involves increasing the difficulty of using the phone and intentionally slowing down user interactions to suppress the desire to use it, such as requiring users to enter random numbers or type self-reflective text [99, 128] and keeping the phone vibrating continuously [93]. The fourth is particularly forceful by directly locking the users' apps or phones for a specific duration [7, 54, 58]. There are concerns about these methods' ability to strike an optimal balance between usability and effectiveness [85].

Internal intervention involves redesigning app interfaces to counteract attention-capturing deceptive designs [87]. For example, increasing user awareness of time spent through reading history labels [11] and specific color change [86], eliminating the addictive design of infinite scroll through removing [79] and adjusting the newsfeed [57, 106, 131], decreasing the guilty pleasure recommendations through using adaptable commitment interface [74] and redesigning search interface [86]. Compared to external intervention, internal intervention can better balance effectiveness and long-term experience [106, 131]. However, these internal methods often require third-party development, as large companies rarely adopt such designs themselves due to financial interests [33]. This necessitates additional development costs and the proposed design is typically tailored for a single app, making it hard to apply broadly.

Therefore, we hope to create a universal external intervention, using the form of reminders to ensure usability while boosting intervention effectiveness through personalized persuasion.

2.4 Persuasion for User Behavior Change and Large Language Models

Persuasion is a psychological approach designed to influence attitudes, beliefs, or behaviors [20]. Language is the most common means of persuasion [108, 114], and leverages facts, emotional appeals, and so on to achieve its goal. Its effectiveness has been shown in multiple fields, such as advertising to encourage consumers to buy a product [13], supporting mental health such as coping with stress [89, 96], and managing physical health such as reducing snacking behavior [50]. For smartphone use intervention, persuasion usually appears as reminders, such as leveraging the user's usage time in a template format [8, 46], or some thought-provoking statements [128]. Prior work has suggested that personalizing content can enhance persuasion effectiveness such as reducing snacks [50, 51]. Also, varying the timing and content of interventions, sometimes even randomly, can improve effectiveness. In contrast, static interventions tend to lose influence over time [60].

The advent of Large Language Models (LLMs), like ChatGPT[94] and PaLM[19], has made vast progress in personalized and diverse content generation. Recent studies have explored various health applications supported by LLMs, such as health information seeking [80, 129, 130], mental health support [61, 65, 126], personal health coaching [88, 122], health education [62], and public health interventions [48]. These applications showcase LLMs' capabilities in knowledge delivery and emotional support. Compared to them, our study further explores LLMs for just-in-time behavior change and intervention, beyond information presentation.

3 MENTAL STATES OF HABITUAL SMARTPHONE USE AND PERSUASION STRATEGIES

To comprehend the mental states of users' smartphone use and guide our intervention system design, we initiated a Wizard-of-Oz (WoZ) study, followed by a semi-structured interview study (Sec. 3.1). We summarized the main takeaways in Sec. 3.2. Based on theories and our findings, we devised four persuasion strategies and their implementation under different mental states (Sec. 3.3).

3.1 Exploratory Wizard-of-Oz & Semi-structured Interview Studies

To identify particular smartphone usage behaviors requiring targeted interventions, we first recruited 12 end-users (6 females and 6 males, aged 18-28) and conducted a 5-day WoZ study in the wild. The findings suggested ideas for persuasion content design. For deeper insights into participants' mental states and concrete intervention design materials, we recruited another group of 10 users (5 females and 5 males, aged 18-29) and conducted a semi-structured interview study[1]. Both studies are approved by the institution's IRB. Our studies focused on young adults who have been reported to have the most severe problematic smartphone use issues [67, 84]. However, we do recognize that our sampling could limit the generalizability of our findings. We discuss this as a limitation in Sec. 8.

3.1.1 Wizard-of-Oz Study. We developed a chatbot system for smartphones that tracks user app activity. First, we asked participants to select apps for intervention, adding them to a blacklist. The chatbot then would send persuasive messages to the participants upon opening a blacklisted app.

To reduce the observation effect, participants were told that it was an automatic chatbot instead of a human [52]. In reality, when participants opened a blacklisted app, a human experimenter would receive an email notification. Based on smartphone usage duration and frequency (see the detection method in Sec. 5), the experimenter designed and delivered persuasive messages. Inspired by existing literature on persuasion design [1, 14, 35], our messages fell into 4 types (see examples in Table 4 in Appendix): (1) usage notice, telling participants their usage data such as the accumulated usage time today and time since last use, (2) practical guidance, asking participants' goals today and suggesting tasks instead of smartphone use, (3) encouragement, praising and cheering participants to keep smartphones away, and (4) deterrent, alarming participants the consequences of using smartphones such as task delay and admonishing them to stop.

Every evening, researchers conducted a brief 15-minute online interview in person with each participant, structured around four questions: (1) What was your overall experience of using the chatbot? How did it change your smartphone usage? (2) Why were

[1]Since the two studies are close in time, we choose a separate set of participants for semi-structured interviews to avoid the impact of intervention in the WoZ study.

you using your phone at a particular time? (3) What were your reactions to the persuasive message, and why? (4) How did you like the persuasive message? How can it be improved? At the study's conclusion, we informed participants that the chatbot was actually operated by a human experimenter at the back end.

All persuasive messages and participants' responses, along with their sending times, are documented. Daily interviews were audio-recorded and transcribed. The recorded data and transcriptions were independently reviewed by three researchers, who coded them based on two main themes: types of smartphone use for question (2) and factors influencing persuasion effectiveness for questions (1), (3), and (4). Subsequently, the researchers convened to discuss the codes until a consensus was reached. Following that, a thorough review of all transcriptions was conducted to ensure the accuracy of the coding.

3.1.2 Semi-structured Interview Study. Our WoZ study provided insights into participants' problematic use behavior and reactions towards persuasion. To obtain a deeper understanding of the user's mental states during phone use, we conducted a semi-structured interview study with another participant group. We asked participants to recollect instances of problematic smartphone use. Our interview started with the question: *"When would you want to use an intervention app to limit your smartphone use?"* We then sought details about the scenario (*e.g.*, time, place, and concurrent activities) and user behaviors and reactions (*e.g.*, usage duration, feelings, and reflections). Next, we asked participants to share their mental states during those instances. We asked questions: *"Why do you use your phone even though you think you should not? What's your mental state behind these reasons?"* We followed the participants' lead during the interview.

All interviews were audio recorded and transcribed. Three researchers independently examined the transcriptions and coded the mental states in different smartphone use cases and contexts. Then they met and discussed the codes until reaching a consensus. To ensure coding accuracy, they went through all transcriptions one more time.

3.2 Main Takeaways about Problematic Smartphone Use

We summarized our main takeaways from the WoZ study and the interview study below. Table 5 in the Appendix summarizes our findings with participants' quotes. Table 1 summarizes the findings about mental states in Takeaway ③ and Takeaway ④.

Takeaway ① Interventions for problematic smartphone use should target habitual usage. Our WoZ study delineated two primary types of smartphone usage: instrumental and habitual, consistent with prior research categorization [76]. We found that only habitual smartphone use warrants intervention. Interventions during instrumental use often led to user dissatisfaction. Notably, relaxation emerged as a crucial form of instrumental use, where participants deliberately used their phones to unwind or reward themselves after intense work or study. Intervening at such times was considered intrusive and inappropriate. The finding aligns with the Dual Systems Theory. In smartphone interactions, instrumental use relies on conscious decision-making (System 2), while habitual use is more instinctive (System 1). Hence, interventions should primarily target habitual use, which is also supported by earlier studies [76].

Takeaway ② The effectiveness of interventions depends on the alignment with users' mental states, personal goals, and contextual information. This is consistent with the literature, suggesting that a shift away from System 2 is due to emotional fluctuations and the absence of defined goals and intentions [78]. We experimented with different persuasive message content during our WoZ study. We found that when we incorporated users' mental states as a factor, which was inferred by their physical contexts and app usage patterns, into generating persuasive message content, participants were more willing to accept the intervention. Furthermore, highlighting users' personal goals enhanced intervention effectiveness. For instance, sending messages like, "*When you find yourself with idle time, consider engaging in meaningful activities such as reading, writing, or drawing*" proved effective when users were in an idle state and had a goal for self-improvement. Our findings are supported by prior studies linking habitual smartphone use to specific mental states [4, 12].

Takeaway ③ Semi-structured interviews revealed three primary mental states connected to habitual smartphone use: *boredom*, *stress*, and *inertia*.

- **Boredom** is an affective state characterized by low arousal and dissatisfaction due to insufficient stimulation [30, 82]. The WoZ and interview studies identified common scenarios leading to boredom: (1) when the task at hand is too simple, lacking a balance between skill and challenge, such as *"doing simple assignments light on cognitive engagement"* (S3)[2] , (2) lack of interest in the current activity, such as *"completing assignments is to relieve a burden, instead of reaching achievement"* (S6), and (3) devoid of any engaging activities during idle moments, such as *"after returning to home"* when is *"not yet time to sleep"* (S8).
- **Stress** refers to cognitive and behavioral reactions to unpredictable and unmanageable stimuli [59]. Participants frequently use smartphones because they experience (1) heightened anxiety when work demands exceed their abilities, such as *"having a challenging bug to locate when doing programming assignment"* (S3) or *"work not progressing well"* (S9) and (2) uncertainty about whether something would have a positive outcome, such as *"not sure if can get a job offer"* (S10). This aligns with increasing evidence that links smartphone use to perceived stress [17, 110, 119].
- **Inertia**, in our context, refers to a psychological resistance that makes users reluctant to change their current activity state. It is similar to the idea illustrated by literature such as emotional inertia [63] or decision-making inertia [2]. Participants stated they commonly used their phones habitually to avoid changing into a new activity state from an idle state, *"checking the phone before starting to do assignment"* (S1) or *"shifting from a relaxed state to a focused state"* (S2). Unlike stress or boredom, inertia does not elicit overt negative emotions but impedes the shift to the next task.

Takeaway ④ Engaging vs. Not Engaging in Activities (Table 1). We further noticed two nuanced categories within mental

[2]This is the serial number of the participant in the semi-structured interview study.

Boredom	
Engaging in Activities	***Not Engaging in Activities***
Users find current tasks boring, lack interest, and struggle to concentrate. This could be due to the task lacking challenges, not being sufficiently engaging, having high repetition, and not aligning with the user's genuine interests and desires.	Users feel bored with daily living in general, lack passion, and have no enthusiasm for engaging in activities. This might be because they have nothing to do, don't know how to pass the time, lack excitement in life, and feel that living is meaningless.
Stress	
Engaging in Activities	***Not Engaging in Activities***
Users feel stressed about current tasks due to challenges posed by the environment, which deplete their resources and result in feelings of tension and unhappiness. This might be due to the abundance, difficulty, and urgency of tasks, causing users to feel anxious, fatigued, and lacking confidence in their abilities, leading to a pessimistic view of the outcomes.	Users feel stressed in the face of daily living and challenges from the environment that deplete their resources, making them feel tense and unhappy. This might be due to setbacks and unexpected events in life that users struggle to adapt to, leading to a pessimistic outlook on the future.
Inertia	
Engaging in Activities	***Not Engaging in Activities***
Users find it difficult to transition from their current state to start the next activity, but without explicit negative emotions. This might be due to procrastination has become a habit, and there's insufficient motivation for the next activity.	Users indulge in idle inertia, but without specific negative emotions. This might be due to idling around has become a habit, and there's no motivation to organize new activities.
Others	
Beyond the scope of the current paper.	

Table 1: Summary of Users' Mental States behind Habitual Smartphone Use.

states. The first category ("engaging in activities") denotes that users have activities to complete while habitually using smartphones. Participants either got distracted from the ongoing boring or stressful activities (*e.g., "I find myself instinctively reaching for my phone in search of mental stimulation when doing simple assignments light on cognitive engagement"* (S3)) or procrastinated to face the upcoming activities (*e.g., "I was reluctant to start handling this challenging work that I scrolled my phone screen anxiously"* (S2)). In contrast, the second category ("not engaging in activities") means users have no schedule or don't know what to do while using smartphones habitually (*e.g., "After getting off work and returning home, I collapse on the sofa and binge-watch Tiktok for one to two hours"* (S9)). Differentiating activity engagement states and combining with the three identified mental states lead to six granular categories of habitual smartphone use, making the persuasion strategies design more situated to users' scenarios. For users engaging in activities, the persuasion not only aims to stop them from smartphone use but also to encourage them to either continue or initiate their activities.

3.3 Persuasion Strategies Design

Based on the takeaways and inspired by the Dual Systems Theory and the Existence, Relatedness, and Growth (ERG) theory, we proposed four distinct persuasion strategies: **Understanding, Comforting, Evoking**, and **Scaffolding Habits**. Developing from Maslow's Hierarchy of Needs, the ERG theory further summarizes human motivation into three levels: (i) the physiological and safety basic needs for existence, (ii) the social needs for feeling related and accepted, and (iii) the need to grow and self-actualize. The theory has been applied to workplaces to increase productivity and job satisfaction [10, 15].

Some of our takeaways align with these theories. For example, **Takeaway ①** aligns with the Dual Systems Theory and suggests that to avoid habitual smartphone use out of instinctive System 1, it is necessary to cultivate enough motivation to maintain conscious but difficult System 2. Moreover, **Takeaway ②** shows that persuasive messages relieving mental states and reminding personal goals are effective, which is consistent with the human motivation of relatedness and growth outlined in the ERG theory.

Accordingly, we map strategies to the relatedness level and growth level of the ERG theory to arouse users' motivation for System 2 (Figure 2). The existence level concerning physiological and safety needs is not included in our theoretical framework. At the level of relatedness, **Understanding** and **Comforting** aim to empathize with users' emotions, offering support and empowering users to manage System 2. At the growth level, **Evoking** reminds users of their personal development goals, and **Scaffolding Habits** guides them in replacing habitual smartphone use with activities conducive to self-fulfillment, thereby turning awareness into action. Then, we map 4 strategies to 3×2 mental states in Figure 3.

3.3.1 Understanding. *Understanding* is a critical strategy to motivate users at the Relatedness level. Past literature suggests that seeking understanding is a coping mechanism [24], and chatbots' empathetic expressions are favored over emotionally neutral advice [71]. Therefore, *Understanding* covers all mental states. This example shows how we integrate understanding into the persuasive content intervention: *"Hi, I know that sometimes you may feel bored and lacking interest. It's okay, this is a very normal feeling. Everyone goes through such times."*

3.3.2 Comforting. *Comforting* aims to comfort users who are experiencing emotional fluctuations, which cause them to shift away from System 2. Prior studies suggest that coping with boredom should focus on meaningfulness [91]. For example, we design a persuasive message to say *"Hey, I know some things might seem a*

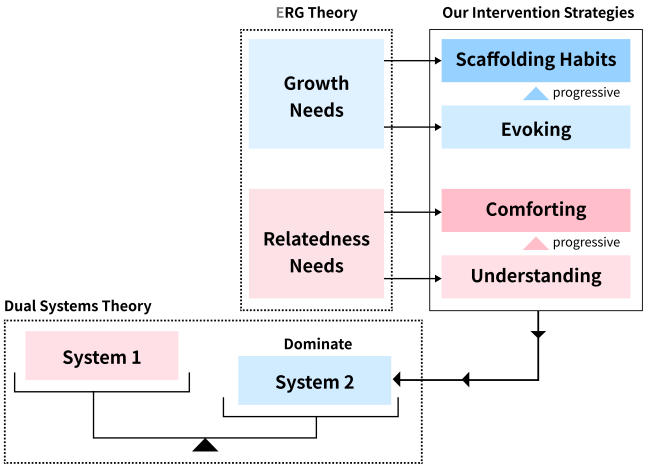


Figure 2: Mapping of Persuasion Strategies to the ERG Theory (Existence, Relatedness, and Growth) and the Dual Systems Theory. According to the Dual Systems Theory, there is a competitive relationship between System 1 (habitual smartphone use) and System 2 (meaningful activity), much like being placed on a scale. To make System 2 heavier than System 1, weights are added—Relatedness needs and Growth needs (the second and third levels of ERG Theory). To support users' Relatedness needs, we design two persuasion strategies: Understanding and Comforting. To motivate Growth needs, we design two other persuasion strategies: Evoking and Scaffolding Habits.

bit boring, but sometimes we need to find the fun in them. Have you ever thought that completing this task would bring you closer to your goals?" In addition, encouragement, humor, acceptance, and wishful thinking [24] can be used to cope with stress by lightening the uncontrollable and unpredictable nature of stressors. For example, *"Don't worry, you have the capability to complete the task! Believe in yourself, and the outcome will pleasantly surprise you."* In the state of inertia without explicit negative emotions, *Comforting* is not employed.

3.3.3 Evoking. Evoking personal goals is a compelling, persuasive technique based on the WoZ study. Literature suggests that goals and values are important for people to sustain System 2 [78] and are closely related to growth motivation. Thus, *Evoking* considers users' goals (*e.g.*, getting high scores in exams, achieving academic success) for designing persuasive strategies. For example, *"Hi! I know you want to play with your phone, but completing tasks is crucial for your IELTS! Keep going, and you're one step closer to a high score!"* This strategy is applied only to scenarios where users are engaging in activities. Goals are arguments used to encourage them to complete or initiate their tasks. In scenarios where users are not engaging in activities, *Scaffolding Habits* assumes the function of *Evoking* by recommending activities that correspond with users' goals, as stated below.

3.3.4 Scaffolding Habits. Last, we encourage users to develop alternative beneficial habits to habitual smartphone use, aligning with their personal value and growth need [78, 101]. By pre-identifying users' preferred habits and considering variables like location and time of habitual smartphone use, we suggest appropriate substitutes to assist users in *Scaffolding Habits*. For example, *"Hi, why not use this moment to memorize vocabulary instead of using your phone? It can help you learn a language and achieve your goals faster!"* *Scaffolding Habits* covers all mental states.

4 MINDSHIFT DESIGN

Building on top of the persuasion strategies we propose in the previous section, we introduce the design of our intervention system: generating persuasive content (what content to intervene with, Sec. 4.1), interaction flow (how to intervene, Sec. 4.2), and intervention timing (when to intervene, Sec. 4.3).

4.1 What Content to Intervene with: LLM-Powered Persuasive Content Generation

We first delved into the importance of context and mental states in persuasion content generation (Sec. 4.1.1). After establishing the significance of context and mental states, we explored how these elements can be intricately integrated into the prompt design (Sec. 4.1.2).

4.1.1 Context and Mental States in Generating Persuasive Content. As suggested in the **Takeaway** ② in Sec. 3.2, the effectiveness of interventions also depends on contextual information. We presented a test case with examples to demonstrate the impact of context and mental states on content generation. In this case, we first constructed a typical college student's context, including time (at late night 00:36 AM), location (in the dorm), and phone usage data (5 mins since the last habitual usage and 10 mins current habitual usage). We then outlined the user's assigned mental state (stressed - engaging in activities), the user's goals (growing research skills, staying healthy) and habits (enjoying outdoor activities), and the corresponding four strategies based on Figure 3.

Differing in context and mental state inclusion, we used GPT-3.5 to generate four sets of persuasive content examples. Table 2 lists the examples of GPT-3.5's outputs, demonstrating that both contextual information and mental state guidance improve content quality. Contextual data empowers GPT to tailor its outputs to the user's current situation, including more poetic sentence phrases such as *"enjoy the night sky outside the window"* or contextual information such as *"You're unstoppable, even at 00:36 AM"*. Adding the mental state guidance enhances GPT's ability to assist users in managing their negative emotions, including empathetic messages such as *"you are not alone"* and encouraging phrases such as *"be closer to completing that great research job"*.

4.1.2 Prompt Design. We constructed four important prompt input factors and fed them to GPT to generate high-quality and persuasive content. As illustrated by purple text in Figure 4, four factors are arranged in a sequence from 1 to 4. We then concatenated them to build a complete prompt as the input to LLMs to generate persuasive content. We tested both GPT-3.5 and GPT-4 and chose to adopt

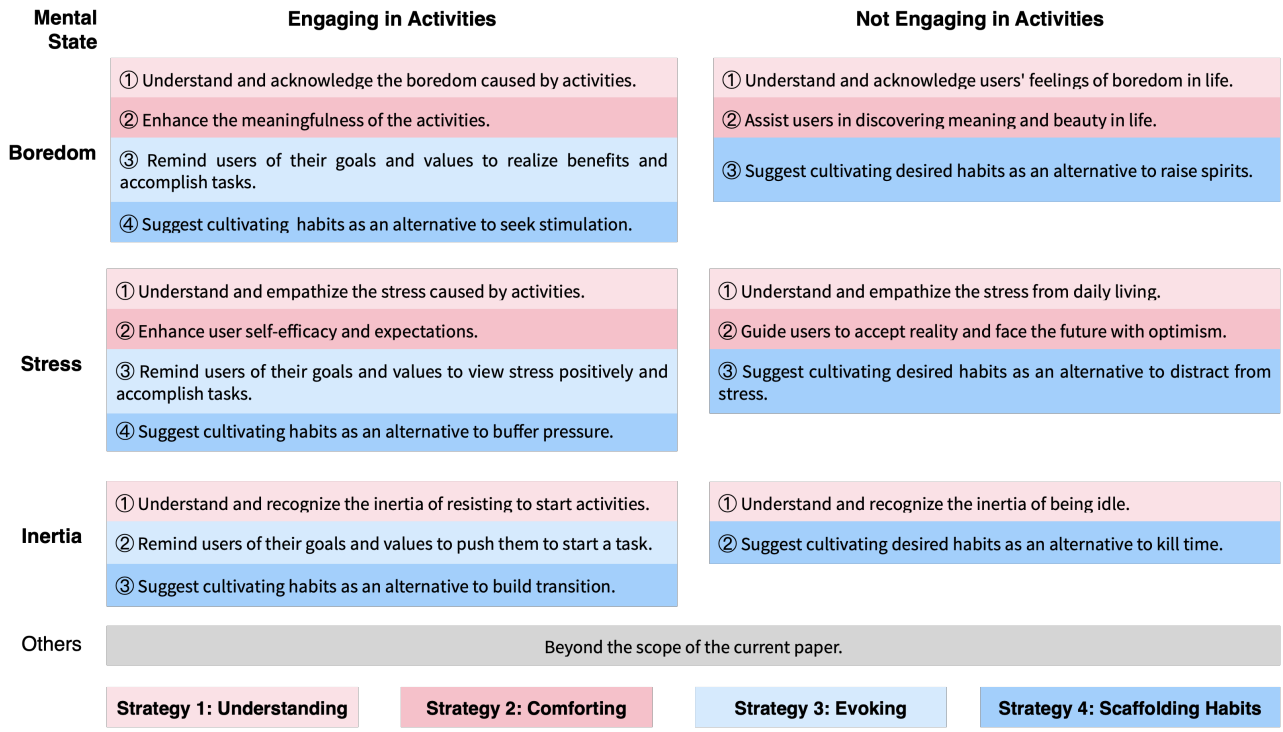


Figure 3: Summary of Persuasion Strategies under Different Mental States. In the table, each color denotes the corresponding persuasion strategy applicable to this scenario.

Table 2: Examples of Generating Persuasive Content with A Test Case Study on Different Context and Mental State Combinations. The letters in parentheses at the end of the sentences indicate the strategies used during sentence generation: U for Understanding, C for Comforting, E for Evoking, S for Scaffolding Habits, and N for No Strategy

	With Mental State	Without Mental State
With Context	1. It's already 00:36, staying up late is stressful, you're not alone, we all feel this way. *(U)* 2. Hey, you've got this! Embrace the challenge with a smile. You're unstoppable, even at 00:36 AM! *(C)* 3. Think about that research article you're about to complete, turn the stress into motivation! *(E)* 4. Try putting down your phone, enjoy the night sky outside the window, relax your eyes. *(S)*	1. It's already 00:36, your phone should probably rest too. *(N)* 2. At this time in the dorm, how about closing your eyes and resting? *(N)* 3. You've already enjoyed 10 minutes of the digital world. *(N)* 4. Just 5 minutes ago, we just said goodbye, meeting again? *(N)*
Without Context	1. Understand your anxiety, it's normal, accept your emotions. *(U)* 2. Relax, each task is a stepping stone to achievement, be optimistic. *(C)* 3. Keep going, you will be closer to completing that great research job. *(E)* 4. How about relaxing in a different way, look far away, let your eyes rest too. *(S)*	1. Friend, your phone might need a rest, and so do your eyes! *(N)* 2. Try putting down your phone, and take a look at the outdoor night views. *(N)* 3. The phone in your hand is not the world, the real fun is around you! *(N)* 4. Every time you put down your phone, it's an opportunity to add points to life! *(N)*

GPT-3.5 as our target LLM to strike a balance between the content generation quality and the speed[3].

[3]GPT-4 introduces a long lag and negatively impacts user experience. Moreover, as we introduce below, the prompts we used as the input for GPT-3.5 include general information that is not individually identifiable. However, we do acknowledge the privacy risk of our method. We will have more discussion in Sec. 8

We lay out the details about how we designed the prompt below:

(1) **Task Setup**: As shown in the left top box in the "Input Prompt to LLM" in Figure 4, Task setup includes <Background> module, providing GPT-3.5 with the global instructions.

(2) **Description of the Current Contexts**: To make each generated content contextually relevant and personalized, it is necessary

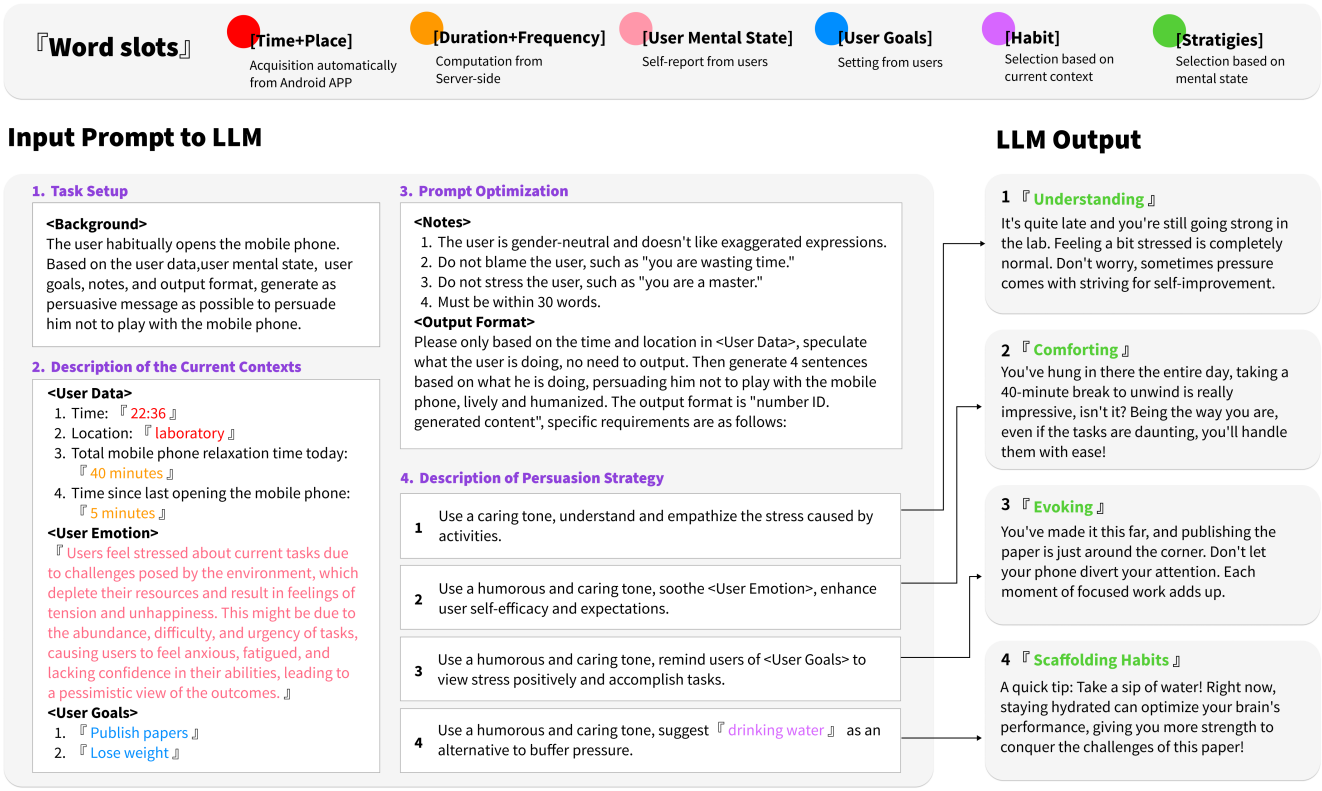


Figure 4: Prompt Templates Used to Generate Persuasive Content with GPT-3.5. The word slots (in the top box) represent different categories of information to be filled in based on the user's current situation. The color indicates the mapping between the slots and the input prompt. The small words under each slot explain the source of the content. The input prompt consists of four parts: (1) Task Setup; (2) Description of the Current Contexts; (3) Prompt Optimization (to improve language quality and reduce harmful content); and (4) Description of Persuasion Strategy (as introduced in Sec. 3 and Figure 3). The LLM output shows the persuasion example of four strategies when the user's mental state is "stressed, engaging in an activity".

to describe the user's current contexts (see the left bottom box in the "Input Prompt to LLM" in Figure 4). This part includes three modules: (a) The <User Data> module describes the user's real-time physical context, including the current time, location, habitual phone usage duration, and the time elapsed since the last habitual phone check. This is collected through phone sensors, see more details in Sec. 5.1. (b) The <User Mental State> module includes users' input from their devices regarding negative emotions and activities. The prompt for this module is selected from the mental state definition (Table 1). This is collected through real-time self-report, see Sec. 4.2. (c) Finally, the <User Goals> module describes what the user values and plays a crucial role in the generation of *Evoking* strategies. Specific user goals are collected during the initialization (Step 0 in Figure 5). This is collected through the initial setup, see Sec. 4.2. The elements mentioned above are represented as word slots (enclosed in brackets in the input prompt), where the users' actual context information can be inserted.

(3) **Prompt Optimization**: To improve the GPT's content quality and effectiveness, we carefully crafted the prompt according to OpenAI's official guidelines [95]. Sometimes LLMs can generate harmful, offensive, or biased texts [38, 132]. We employed an iterative prompt design process to ensure that the persuasive content is appropriate. Initially, one researcher created initial prompts and generated content using GPT-3.5, for six mental states (Figure 3) with five iterations each. Subsequently, two other researchers rated the satisfaction level of the generated content (scoring from 1 to 5), iterating until average satisfaction exceeded 4 to create content that is concise, appropriate, and engaging, while also aligned with our persuasive strategy. Through this process, we addressed some issues with LLM-generated content, such as gender-biased expressions and deviations from human preferences (like being overly exaggerated or stressful) by adding an additional <Notes> section to instruct LLMs' generation. We also took steps to prevent hallucinations by avoiding certain real-world fact-related prompt statements that LLMs can easily make mistakes. By specifying the need to consider current activities in the <output format>, we've made the outputs more contextually relevant. We note that this process cannot fully address the ethical concerns, which we further discuss in Sec. 8.3.3. Our final optimization prompt, shown in the middle top box in Figure 4, includes two modules: (a) The

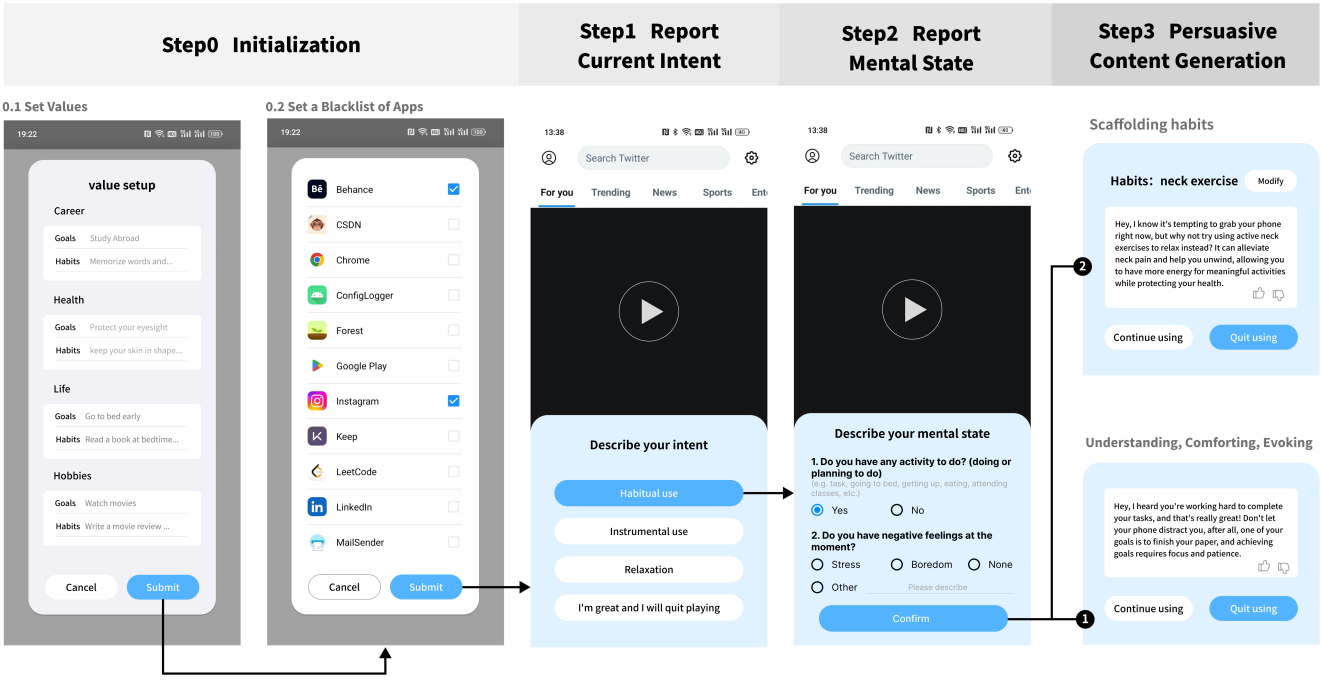


Figure 5: Interaction Flow. Users first complete the global settings for their value and app list (Step 0). When opening a black-listed app, users need to first self-report their phone usage intent (Step 1). If the intent is habitual use, the app asks them to report their current mental state (Step 2). After that, a corresponding persuasion shows (Step 3).

<Notes> module clarifies restrictions; (b) The <Output Format> module guarantees the correct output format.

(4) **Description of Persuasion Strategy**: To ensure each generated content aligned with our proposed strategies, we provide a short description of each strategy based on our design in Sec. 3 and Figure 3, as indicated by the middle bottom box in Figure 4. Based on users' in-the-moment mental states, we will select the corresponding strategies. Note that for the *scaffolding habits* strategy, we also input a habit selected from users' initialization.

4.2 How to Intervene: Interaction Flow

This section outlines the design of the interaction flow in our application. Our interaction process needs to achieve three functions: (1) collect users' phone usage intent to identify habitual use, (2) obtain necessary information for prompt construction, and (3) display the generated content.

Following Takeaway ① in Sec. 3.2, interventions should be targeted at habitual usage. Since automatic detection methods are unreliable (discussed further in Sec. 8), we ask users to self-report, and the system only triggers intervention when the user reports habitual use. In our prompt design in Sec. 4.1.2, we need two categories of information from users: their goals and habits, and their mental state. Goals and habits tend to be stable. So we integrated them into the app's settings page, and users could adjust them as needed. Mental states, however, are more dynamic. Therefore, we captured them through participants' self-reporting. In summary, each intervention episode includes three steps: (Step 1) the user reports their usage intent, (Step 2) the mental state, and (Step 3) the corresponding generated content is displayed, as shown in Fig. 5. Our final design is as follows:

Step 0: Initialization. When initiating the MindShift app, we ask users to complete two global settings. The first is to set their values in four categories: career, health, life, and hobbies, detailing their goals and habits in each. This process serves two purposes: first, it provides the necessary goals for our *Evoking* strategy; second, it enables the creation of personalized habits in the *Scaffolding Habits* strategy. The second is to set a blacklist of apps. Launching apps from this list will trigger intervention.

Step 1: Intent report. Interventions are only necessary when users habitually use their phones (**Takeaway ①**). To collect users' intents, we employ a self-reporting approach. Every time a black-listed app is first opened during an unlock session, users choose from three options: "Habitual use", "Instrumental use", and "Relaxation". Only when users select "Habitual use", the intervention will proceed. Automatically detecting all use is beyond the scope of this paper. We envision that future work can automate this process, as discussed in Sec. 8.

Step 2: Current mental state report. The mental state is a key factor that triggers users' habitual phone use. Unlike physical context, detecting mental state is challenging due to the lack of mature techniques. Therefore, we ask users to self-report. Based on the mental states listed in Sec. 3.2 and **Takeaway ③ & ④**, we propose two single-choice questions for users to report their mental state: (1) whether they are engaged in activities ("Yes" or

"No"), and (2) whether they currently have any negative feelings, including options of "Stress", "Boredom", "None" (*i.e.*, inertia), or "Other Negative Feelings". The "Other Negative Feelings" option, with an adjacent text box for specifics, covers unlisted emotions. We tested these questions for reliability and validity [103]. To ensure validity (*i.e.*, accurate reflection of mental states), three psychology experts verified the alignment of our questions with mental state coding in Sec. 3. To ensure reliability (answer consistency), we asked three pilot study participants to respond to situational mental state descriptions with two single-choice questions, achieving uniform responses.[4]

Step 3: Persuasion. Based on our intervention content design in Figure 4, we leverage the power of GPT-3.5 to generate persuasive messages. The messages are displayed in a pop-up window, where users can choose to either quit the app or continue using it (the bottom of Step 3 in Figure 5). Additionally, users can also provide optional feedback by giving a thumbs up or down.

Furthermore, for the *Scaffolding Habits* strategy, it is essential to link users' own habits to specific use contexts. Therefore, we implement a user participation mechanism, adding an additional habit item along with an edit button (the upper interface of Step 3 in Figure 5). The habit item represents a system-generated suggestion based on the user's current context and initial settings in Step 0. Users can edit and update their desired habits. Once submitted, the modified habit will be recommended the next time when users are in the same context.

4.3 When to Intervene: Intervention Trigger Mechanism

We consider the user's intent of use in the intervention trigger rule (Step 1 of Sec. 4.2). Specifically, interventions will be triggered when a user's self-report intent is habitual use.

As in Figure 3, each mental state allows for multiple persuasion strategies. We devise a simple procedure. After determining the mental state and narrowing down the specific strategies, we first randomly sample one strategy and generate an intervention message. Then, we loop over other strategies and show new strategy messages every two minutes until the users leave the app. After looping over all appropriate strategies under this mental state, the intervention will stop. The usage duration is calculated based on the total usage time of a single blacklisted app during one unlocking session. The two-minute interval setting is derived from the statistical analysis in the WoZ study, where 90% of users spent less than 5 minutes on a blacklisted app in a single session. We thus set the interval between interventions as two minutes as a convenient delay to facilitate the exploration of different strategies, as further supported by the analysis in Sec. 7.2.2.

Users only need to report their habitual use and mental states once (*i.e.*, Step 1 and 2 in Figure 5) when they open a specific app during each screen unlock session. This design is based on three considerations: (1) The initial mental state when opening an app is crucial, as it triggers habitual use (Sec. 3.2). (2) As reflected in our pilot study in Sec. 4.2, multiple reports during app switching can be annoying and negatively affect user experience. (3) Some previous studies suggest that users' stress level tends to be retained even with coping techniques [27]. We assume this also applies to other mental states, so most users' mental state remains stable during one session (90% was less than 5 minutes), as further supported by the analysis in Sec. 7.2.1. We also discuss future potential ways to enhance accuracy in Sec. 8.3.2 & 8.4.

[4]We plan to conduct more comprehensive validity and reliability testing in future work, as discussed in Sec. 8.

5 SYSTEM IMPLEMENTATION

We built an Android application to instantiate our design. Our system consists of a client and a server.

5.1 Client-Side Implementation

The client-side is an Android app that implements all the features of our design. We use accessibility services to detect the opening and closing of apps on the phone. The client is also responsible for collecting and uploading data to the server, including screen's off and unlock status, application name, application opening and closing times, and location data obtained through the Amap API [3]. To prevent the accidental killing of the accessibility service and ensure compliance, our app includes a background service checking the accessibility service status every 5 minutes. If it detects that the service is terminated, the app informs the server to email a researcher, who then reminds the user to reactivate the service, ensuring data integrity.

5.2 Server-Side Implementation

The server side is responsible for generating persuasive content through four key tasks, ensuring that the intervention is personalized, contextually relevant, and delivered in a timely manner.

(1) **User Data Computation**: The server processes user data from client-uploaded app data for use in word slots, including the phone's total habitual use time and last habitual opening time.

(2) **Habit Selection**: Next, the server selects a habit mostly matched with the current user's mental state, location (*i.e.*, the specific building), and time (*i.e.*, the hour of the day) from the users' initialization (*i.e.*, Step 1 in Figure 5). To reinforce the habit-context link, unless users thumb down or modify it, the same habit will be recommended in the same context. More details can be seen in Step 4 of 4.2.

(3) **Strategy Counterbalance**: To balance the frequency of each strategy across mental states, the server counterbalances strategy order in the prompt in Sec. 4.2.

(4) **Content Generation**: After obtaining the user contexts, habits, and persuasion strategies, the server uses the OpenAI GPT-3.5 API to generate persuasive content. We adopted a streaming, character-by-character generation approach, allowing the persuasive content to start being displayed within 2 seconds.

6 FIELD EXPERIMENT

To evaluate the effectiveness of MindShift, we conducted a 5-week field experiment. We introduce experimental design (Sec. 6.1 & 6.2), participant recruitment (Sec. 6.3), and experiment procedure (Sec. 6.4).

Table 3: Comparison of Three Intervention Methods

Intervention Methods	Characteristics		
	Intent Report	LLM-powered Persuasion	Mental-States-Based Persuasion Strategies
Baseline	✓		
MindShift-Simple	✓	✓	
MindShift	✓	✓	✓

6.1 Baseline and MindShift-Simple Intervention Methods

As *MindShift* is one of the first persuasion intervention systems leveraging an LLM to generate dynamic persuasion content, there are no comparable systems other than the traditional persuasion techniques. We compared *MindShift* against a persuasive reminder baseline, one of the most commonly adopted intervention methods in commercial apps [8, 46]. To ensure the fairness of the comparison, the baseline is designed to be the same as the intent report step, as illustrated in Step 1 in Figure 5. Specifically, it only requires users to report the intent the first time they open the blacklist app after unlocking. It can also be used to collect the proportion of users' initial intents, facilitating the analysis of the intervention effect of *MindShift*. We name this intervention as *Baseline*.

Moreover, in order to evaluate the effectiveness of the mental states and persuasion strategies proposed in Sec. 3, we further designed a simplified version, *MindShift-Simple*, by removing the mental states and persuasion strategies from the prompt design. Specifically, in the prompt design (Figure 4), we retained only <User Data> in the (2) Description of the current context and removed the (4) Description of the persuasion strategy. Meanwhile, the last sentence in <Output Format> was changed to generate four sentences at once. We kept other setups consistent, ensuring that both versions' language features (such as both tone styles are humorous and caring) are as consistent as possible. Examples of content generation in *MindShift-Simple* are as shown in Table 2 under 'With Context' and 'Without Strategy'.

In total, we have three intervention methods to compare: *Baseline*, *MindShift-Simple*, and *MindShift*. Table 3 shows the comparison. Figure 12 in Appendix further shows their interaction flow.

6.2 Experiment Design

We adopted a within-subject design, with intervention techniques as independent variables (*Baseline*, *MindShift-Simple*, and *MindShift*). We designed a 5-week field experiment. To measure users' everyday phone usage behavior, the first week is set as the *Baseline* stage, followed by two weeks of one *MindShift* version and another two weeks of the other version. We counter-balanced the order of *MindShift-Simple* and *MindShift*.

Our evaluation metrics include various aspects: (1) Intervention acceptance rate. We measure the percentage of times users accept the intervention and quit the blacklist app use when interventions are shown. (2) Intervention thumb-up rate. For *MindShift-Simple* and *MindShift* that show persuasion content, users can provide feedback by thumb-up or thumb-down (step 4 in Figure 5). (3) App usage behavior, which includes both app opening frequency and usage duration. (4) Subjective reports. At the beginning of the study and at the end of each intervention session, we distributed the Smartphone Addiction Scale (SAS) [64] and the self-efficacy scale [111]. In addition, at the end of the study, we also conducted a brief semi-structured interview to gather user experiences and feedback on our intervention techniques. These metrics cover both the objective and subjective measures of the interventions.

6.3 Participants

We sent out recruitment material on social media platforms. We included a screening survey aiming to identify potential participants who showed signs of smartphone addiction and the willing to reduce their smartphone use. Specifically, besides basic demographics, we included four questions selected from the SAS and self-efficacy questionnaires, questions about the willingness to reduce smartphone use, the extent of habitual phone use, future plans for the next five weeks, and a screenshot of phone usage time of the last week. We excluded users (1) without signs of smartphone addiction (SAS sub-score < 15), or (2) unwilling to reduce smartphone use, or (3) less than 20 hours of weekly phone use, or (4) having special plans such as long-term travel in the next five weeks (which may shift their phone usage patterns).

We received a total of 42 responses. We recruited 31 participants after the screening process. 6 participants voluntarily dropped out during the study. For the remaining participants, we divided them into groups according to counterbalanced intervention orders. We conducted a Kruskal-Wallis test analysis to ensure that groups had no significant difference in the SAS scores and self-efficacy scores. In the end, 25 participants completed the entire study (females=13, males=12, age=22±2 years), including 17 undergraduates, 5 graduate students, and 3 professionals.

6.4 Experiment Procedure

After all the participants signed the consent form, we held a 20-minute onboarding session online to familiarize participants with the research process and introduce the Android application. We explained in detail the meanings of each selection in the intent report interfaces (Step 2 in Figure 5). After the meeting, participants filled out the first SAS and self-efficacy questionnaires. The app was then deployed for a 5-week field experiment.

Before users started using one of two versions of *MindShift* (*MindShift* and *MindShift-Simple*), we provided participants with a tutorial explaining the three mental states (*i.e.*, boredom, stress, and inertia) that they need to report. To confirm their understanding,

participants were asked to take a brief test on the understanding of mental states until they reached a score of 90%. This ensured that they had an accurate and consistent understanding of the mental states, which in turn improved the accuracy and reliability of the reported data. At the end of the experiment, participants received a compensation of $50 for their time.

7 RESULTS

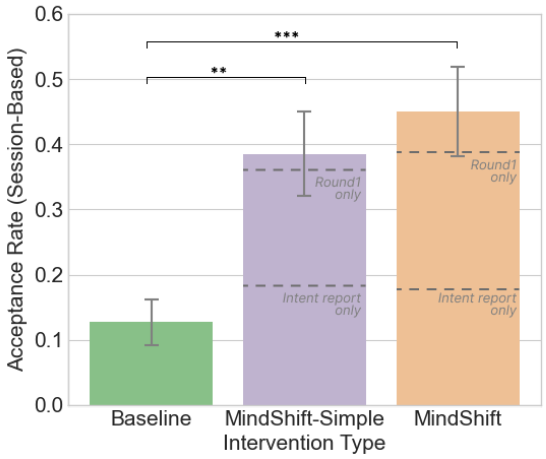
During the five-week study, we collected 50,815 minutes of restricted app usage duration, and 54,467 restricted app opening events (7539, 23,769, 21,994, and 835 for habitual use, instrumental use, relax, and quit). We conducted statistical tests on the quantitative data collected from the app and scale scores to measure differences. For qualitative data from exit interviews, we conducted thematic coding to extract key insights.

7.1 Intervention Acceptance Rate

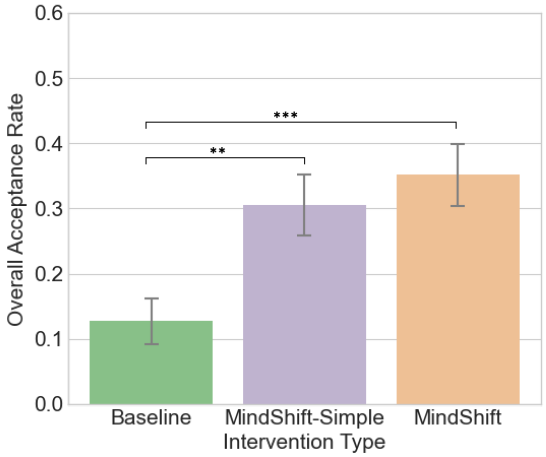
The effectiveness of a persuasion strategy is directly measured by the rate of successful prevention of user engagement with the targeted application (*i.e.*, intervention acceptance rate). We compared the overall acceptance rate (Sec. 7.1.1), the acceptance rate for generated persuasion content (Sec. 7.1.2), and thumb-up rate (Sec. 7.1.3). Moreover, we also conducted a detailed analysis of the acceptance rate across strategies (Sec. 7.1.4), mental states, and activities (Sec. 7.1.5).

7.1.1 Overall Acceptance Rate. ***MindShift* and *MindShift-Simple* increase the overall acceptance rate significantly and *MindShift* achieves best.** We assess the overall acceptance rate using two methods. The first "session-based" rate means among total app visits (excluding instrumental uses and relaxation), how many times users quit during the intervention (including intent report and persuasion content [5]). Figure 6a shows that *MindShift* (45.1±34.3%) achieves higher acceptance than *MindShift-Simple* (38.5±32.2%) and *Baseline* (12.7±17.4%). Significance is observed in a Friedman test ($\chi^2(2) = 16.64$, $p < .001$). Three post hoc Wilcoxon signed-rank tests, corrected with Holm's sequential Bonferroni procedure, indicate that *MindShift* vs. *Baseline* ($V = 31$, $p < .001$) and *MindShift-Simple* vs. *Baseline* ($V = 59$, $p < .01$) are significantly different, while *MindShift* vs. *MindShift-Simple* is not ($V = 126$, *n.s.*).

Considering that the *Baseline* doesn't trigger subsequent interventions like the *Mindshift* and *Mindshift-simple*, we also compare the acceptance rates of the first round in particular. Both *MindShift* and *MindShift-Simple* initiate persuasion immediately after participants report their intents, so we include this initial persuasion in calculating their first-round acceptance rates. Additionally, we analyze the acceptance rate of the intent report to distinguish its effectiveness among the three intervention techniques. As shown in the upper dashed lines in Figure 6a, the first-round acceptance rates for *MindShift* (38.7%±24.9%) and *MindShift-Simple* (36.2%±25.8%) are still significantly higher than the *Baseline* (12.7%±17.4%,$\chi^2(2) = 15.56$, $p < .001$). Post hoc tests show significant differences in *MindShift* vs. *Baseline* ($V = 43$, $p < .01$), and *MindShift-Simple* vs. *Baseline* ($V = 32$, $p < .001$), but not in *MindShift* vs. *MindShift-Simple* ($V = 134$, *n.s.*). The lower dashed lines in Figure 6a indicate that the acceptance rates when considering only reporting intent are still higher for *MindShift* (17.9%±17.5%) and *MindShift-Simple* (18.3%±18.2%) compared to the *Baseline* (12.7%±17.4%). However, a Friedman test ($\chi^2(2) = 5.59$, $p < .1$) does not show significance, suggesting that intent report has no difference among the three intervention techniques.



(a) Overall Acceptance Rate (Session-based)

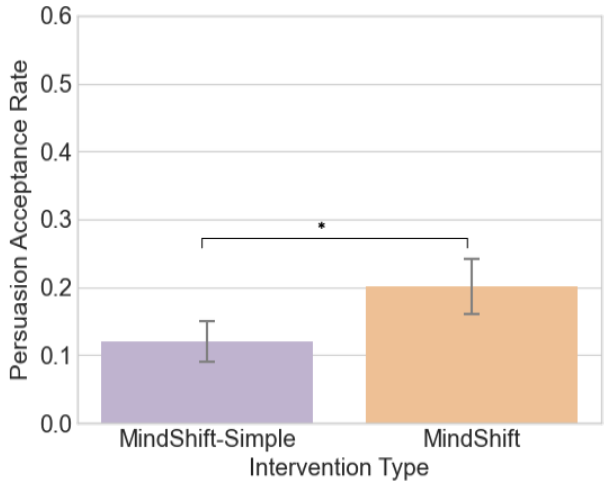


(b) Overall Acceptance Rate (Pop-up-based)

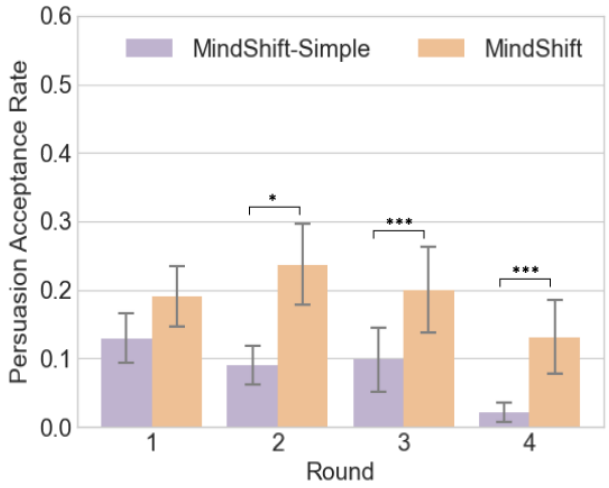
Figure 6: Overall Acceptance Rate

Following the session-based method, we also investigate the pop-up-based acceptance rate, which equals the total number of quit times divided by the total number of intervention pop-ups (i.e., each round is counted as a pop-up). Figure 6b shows the comparison, *MindShift* (35.2±24.1%) still has a higher acceptance than *MindShift-Simple* (30.5±23.6%) and *Baseline* (12.7±17.4%, $\chi^2(2) = 13.69$, $p < .01$). Post hoc tests indicate significance for *MindShift* vs. *Baseline* ($V =$

[5] Users click the "I am great and I will quit playing" in Step 1 or "Quit using" in Step 3 in Fig 5

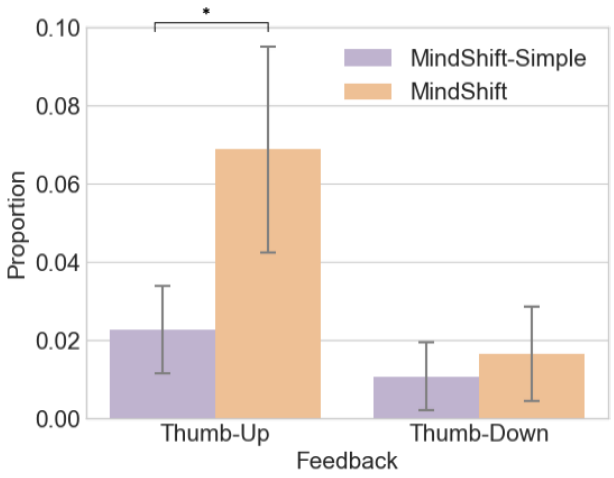


(a) Overall Persuasion Acceptance Rate

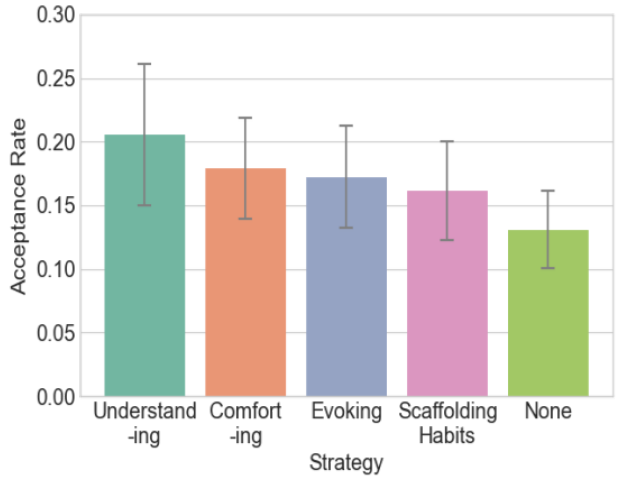


(b) Persuasion Acceptance Rate Grouped by Round

Figure 7: Persuasion Acceptance Rate



(a) Feedback Proportion



(b) Acceptance Rate Grouped by Persuasion Strategies

Figure 8: Feedback and Strategy Acceptance Rate

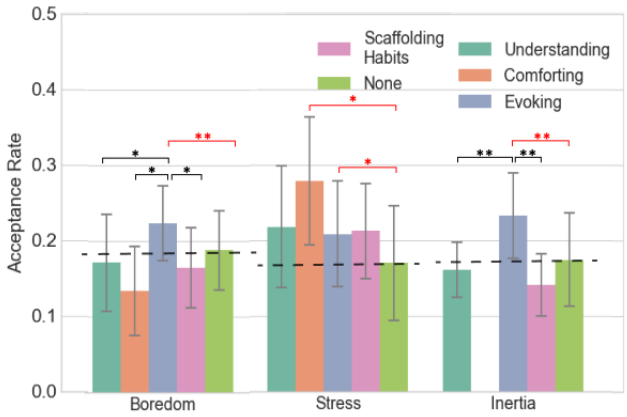
52, $p < .01$) and *MindShift-Simple* vs. *Baseline* ($V = 38$, $p < .001$), but not for *MindShift* vs. *MindShift-Simple* ($V = 129$, *n.s.*). Subsequent analyses are all based on pop-ups.

7.1.2 Acceptance Rate for Generated Persuasion Content. ***MindShift* has a significantly higher persuasion acceptance rate than *MindShift-Simple*.** The overall acceptance rate includes two parts: exiting when reporting intent and exiting after seeing the persuasion content. To narrow down the comparison between *MindShift* and *MindShift-Simple*, we exclude the intent report stage and focus on the acceptance rate during the persuasion stage, as they differ only in the persuasive content. As shown in Figure 7, *MindShift* achieves higher persuasion acceptance (20.1±20.2%) than *MindShift-Simple* (12.0±15.0%) and a paired-samples t-test shows that *MindShift* was statistically significantly higher (t = -2.21, p<0.05).

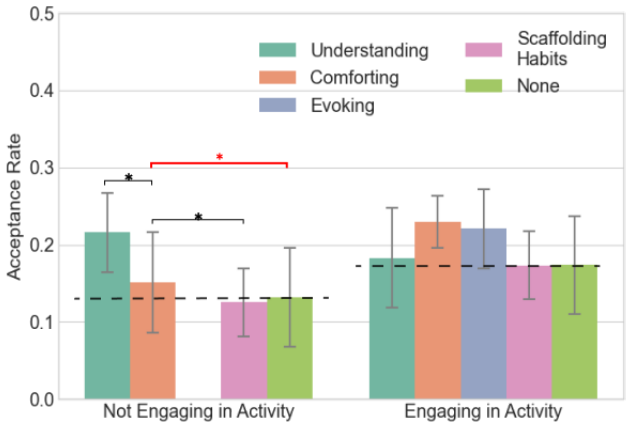
Moreover, as we introduce in Sec. 4.3, every persuasion intervention could consist of 1 to 4 rounds, depending on which mental state participants are in and which stage participants leave the app. Therefore, we further compare the persuasion acceptance rates between *MindShift* and *MindShift-Simple* across different persuasion rounds. The results show that *MindShift* outperforms *MindShift-Simple* at each round [6](Δ_{Round1}=6%, Δ_{Round2}=14.7%, Δ_{Round3}=10.2%, Δ_{Round4}=11%) as indicated in Figure 7b. We conduct a paired-samples t-test between the two intervention techniques in each round. Results indicate that, except for the first round, rounds 2 (p<0.05), 3 (p<0.001), and 4 (p<0.001) all exhibit that *MindShift* achieves significantly higher acceptance rate. This trend suggests that as the number of interventions increases, *MindShift*'s advantage becomes more pronounced, highlighting *MindShift*'s robustness and effectiveness in maintaining high acceptance rates.

7.1.3 Thumb-up Rate of Interventions. ***MindShift* has a significantly higher thumb-up rate.** Users can give feedback in the

[6]Round 1 here only contains persuasion and excludes the intent report.



(a) Strategy Acceptance Rate Grouped by Mental State. Significant differences compared to no strategy (green bar) are highlighted in red, while differences among strategies are indicated in black.



(b) Strategy Acceptance Rate Grouped by Activity. The same annotation method as grouped by mental state.

Figure 9: Strategy Acceptance Rate Grouped by Different Mental States

persuasion interface. As depicted in Figure 8a, 6.8% of interventions in *MindShift* receives thumb-up while *MindShift-Simple* receives only 2.2%. A paired-samples t-test shows that significant differences (p<0.05) are observed for the thumb-up rate, but no significant differences (p=0.22) for the thumb-down rate between the two techniques. This indicates that *MindShift* aligns better with users' preferences.

7.1.4 Acceptance Rate across Different Strategies. We design four persuasion strategies in *MindShift* whereas *MindShift-Simple* does not incorporate any specific strategies, so we further compare the persuasion acceptance rates across different strategies. Figure 8b shows that all strategies we design outperform *MindShift-Simple* ($\Delta_{Understanding}$=8.9%, $\Delta_{Comforting}$=3.3%, $\Delta_{Evoking}$=10.3%, $\Delta_{Scaffolding\ Habits}$=4.9%) but the differences are not statistically significant.

7.1.5 Strategy Acceptance Rate across Different Mental States and Activities. As we show in Figure 3, each mental state has a different strategy mapping. Therefore, we also seek to derive insights regarding which strategies are most effective for users under different mental states (Figure 9a) and activities (Figure 9b). Friedman test and post hoc Wilcoxon signed-rank tests are employed to investigate the influence of different strategies on acceptance rate.

For mental states: under the mental state of *"Boredom"* and *"Inertia"*, *Evoking* is significantly more effective ($\Delta_{Boredom}$=3.6%, $\Delta_{Inertia}$=5.8%) than *MindShift-Simple* (ps < 0.01); under the mental state of *"Stress"*, *Comforting* and *Evoking* show a trend toward significance compared with *MindShift-Simple* (Δ = 10.8%, p<0.1 and Δ = 3.8%, p<0.1 respectively).

For activity levels: when not engaging in activity, *Comforting* is significantly more effective than *MindShift-Simple* (Δ=5.6%, p < 0.05); when engaging in activity, there are no significant differences in all the strategies compared to *MindShift-Simple*.

7.2 App Usage Behavior

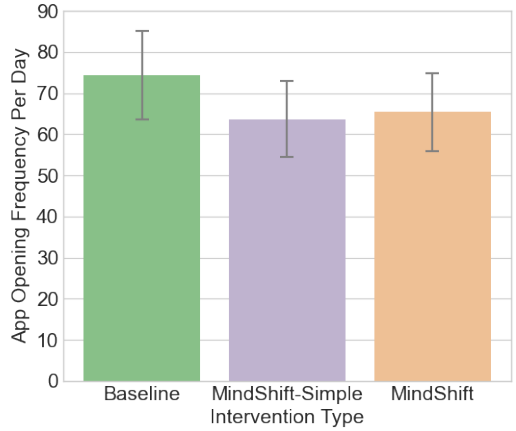
We then investigate the influence of the intervention on participants' app usage behavior. Overall, participants have less app usage frequency and duration when using *MindShift* and *MindShift-Simple*, especially in habitual usage.

7.2.1 Overall Usage Behavior. We count the number of app opening attempts for restricted apps. Figure 10a presents the opening frequency (daily open count) under three intervention techniques. *MindShift-Simple* (63.6±9.2) and *MindShift* (65.3±9.5) have lower opening frequency than *Baseline* (74.3±9.7). Compared to *Baseline*, *MindShift* reduces by 12.1% usage duration while *MindShift-Simple* reduces by 14.4%. However, a Friedman test does not show significance.

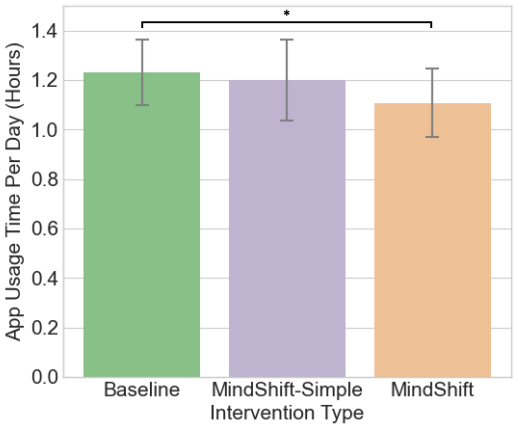
We also measure restricted app usage duration, another important factor for phone overuse. As can be seen from Figure 10b, participants have the lowest app usage duration in *MindShift* (1.11 ±0.7 hours) compared to the *Baseline* (1.23±0.7 hours) and *MindShift-Simple* (1.20 ± 0.8 hours). Compared to *Baseline*, *MindShift* reduces by 9.8% usage duration while *MindShift-Simple* reduces by 2.4%. We conduct a Friedman test and find a significant difference among different techniques (p<0.05). A post-hoc Wilcoxon test shows that *MindShift* has a trend of declining compared to *Baseline*, with marginal significance (p<0.1).

To validate the assumption that the intent and mental state remain stable in one unlock session across different apps in Sec. 4.3, we analyze the duration of users' intent and mental states. The results show that the median duration of a mental state is 5 hours (third quartile 14.5 hours). The median duration for an intent (changing from habitual use to other intents) is 37 minutes (third quartile 60 minutes). This validates our hypothesis that intent and mental state are stable during one habitual usage session.

7.2.2 Habitual Usage Behavior. ***MindShift* and *MindShift-Simple* significantly reduce habitual app usage duration and frequency.** The focus of our intervention is habitual use, so we investigate the changes in habitual usage behavior. Results show that *MindShift* and *MindShift-Simple* can both significantly reduce habitual use. The app visit frequency and duration of habitual usage cases also decrease significantly during the two versions of *MindShift* compared to *Baseline* ($\Delta_{MindShift\text{-}Simple}$ equaled 80.6% for

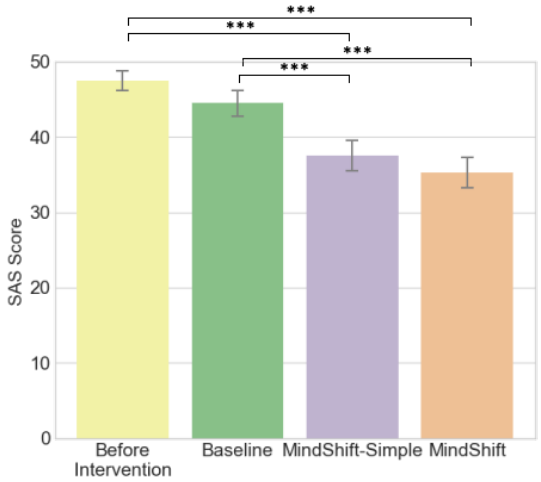


(a) Total App Opening Frequency

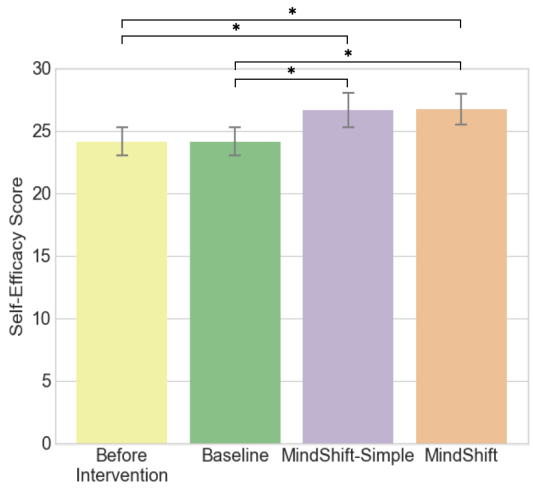


(b) Total Phone Usage (Hours) per Day

Figure 10: App Usage Behavior



(a) SAS Score



(b) Self-efficacy Score

Figure 11: Subjective Scales Report

visit frequency, 84.4% for usage duration, and 6.8% for habitual use proportion, $ps < 0.001$; $\Delta_{MindShift}$ equaled 77.3% for visit frequency, 80% for usage duration and 6.8% for habitual use proportion, $ps < 0.001$).

To confirm the suitability of the 2-minute intervention interval, we analyze data on users' habitual usage duration during the *Baseline* phase (unaffected by subsequent interventions). Analysis shows that 75% of users spent about 4 minutes in a single habitual use, supporting our design choice of the 2-min intervention interval.

7.3 Subjective Report

We further analyze on user-reported SAS and self-efficacy scale results. When using *MindShift* and *MindShift-Simple*, participants experience a significant decrease in SAS score and a significant increase in self-efficacy score, but they see no change when using *Baseline*. We summarize the results as follows.

7.3.1 Decrease of SAS Score. ***MindShift* and *MindShift-Simple* decrease SAS score significantly and *MindShift* achieves best**. Figure 11a shows the results of the SAS scores during the intervention stages. *MindShift* exhibits the lowest SAS scores (35.2±10.2), followed by *MindShift-Simple* in the second position (37.6±10.1), with the *Baseline* intervention ranking the third (44.5±8.7) and the initial ranking the last (47.5±6.7). This indicates that *MindShift* and *MindShift-Simple* reduce SAS scores by 34.7 and 25.8%. The results of a Friedman test show a significant difference ($p<0.001$). Post-hoc Wilcoxon tests show that both *MindShift* and *MindShift-Simple* are significantly lower than the *Baseline* and initial scores (all $ps<0.01$). This suggests that the two versions of *MindShift* have the potential to fundamentally alter individuals' mobile phone usage behavior.

7.3.2 Increase of Self-efficacy Score. ***MindShift* and *MindShift-Simple* increase Self-efficacy score significantly**. Figure 11b shows the results of the self-efficacy scores during the intervention stages. *MindShift* (26.7±6.1) and *MindShift-Simple* (26.6±6.8) exhibit

higher scores compared to the *Baseline* intervention (24.1±5.6) and the initial (24.1±5.6). This indicates that *MindShift* and *MindShift-Simple* increase self-efficacy scores by 10.7% and 10.4%. Friedman test shows a significant difference (p<0.001). The post-hoc Wilcoxon test shows that both *MindShift* and *MindShift-Simple* are significantly higher than the *Baseline* and initial scores (all ps<0.05). This indicates that persuasive techniques have the potential to enhance individuals' self-efficacy, which can result in overcoming excessive mobile phone usage. In contrast, conventional reminders may not achieve this goal.

7.3.3 Subjective Comments. During the exit interview, participants generally had a positive experience. One participant said, *"I can feel that my recent dependency on the phone has decreased"* (P17). One participant felt using the app shifted them to self-improvement tasks instead of mindlessly usage, "*Now, when I have nothing to do, I tend to do other things, like learning vocabulary, instead of aimlessly browsing my phone"* (P10). Another participant was willing to use it longer, *"I'm a little sad with the disappearance of pop-ups after the experiment. If possible, I would like to keep using it"* (P3). Additionally, participants had positive comments on *MindShift* which includes the mental states factor, "*I feel that its suggestions align well with my emotional state at that time.*" (P25). They also valued the "*Understanding*" and "*Comforting*" strategies, saying "*It tells me that I'm not the only one experiencing these painful emotions, which helps me feel better*" (P22).

Despite the majority of positive comments, a small number of participants expressed their dissatisfaction with *MindShift*. Some participants found the persuasive message to be "*a bit stiff and templated*" (P4, P6), and they believed that "*they would develop tolerance as they repeatedly use"* (P20). Some participants also mentioned privacy concerns. P15 mentioned that *MindShift-Simple*'s ability to capture time and location made her uncomfortable. Moreover, participants' preference for linguistic characteristics is highly personal. Some felt harsh ones were more useful, *"Gentle tone doesn't work for me, I wish the words could be harsher"* (P8). Some preferred data proof than pure textual reasoning, *"It's intuitive to tell me how long I have used my phone today directly. The number is very eye-catching"* (P9). This suggests the future direction of personalized persuasive content design. We have more discussion in Sec. 8.2.

7.4 Summary of Results

Overall, two versions of *MindShift* show significantly higher acceptance rates compared to the *Baseline*. *MindShift* has the highest acceptance (45.1% for session-based and 35.2% for pop-up-based) and thumb-up rates (6.8%), and it statistically significantly outperforms *MindShift-Simple* in both acceptance rates of generated persuasion content (8.1%) and on a per-round basis (6-14.7%). Furthermore, there exist strategies that significantly outperform *MindShift-Simple* (3.6-10.8%) in every mental state, suggesting that the strategies we design are meaningful. Furthermore, *MindShift* and *MindShift-Simple* lead to a decrease in overall app opening frequency (12.1-14.4%) and usage duration (9.8-2.4%) and are significantly effective in reducing habitual use. *MindShift* and *MindShift-Simple* also reduce SAS scores by 34.7-25.8% and increase self-efficacy scores by 10.7-10.4% statistically significantly while *Baseline* does not. This suggests that *MindShift* has the potential to profoundly transform human behavior with enduring effects. Finally, users' subjective comments also confirm a perceived reduction in smartphone dependency and an inclination to continue using *MindShift*.

8 DISCUSSION

In this section, we discuss *MindShift*'s novelty in contrast to previous intervention techniques (Sec 8.1), future work (Sec 8.2), the potential of leveraging LLMs for behavior change (Sec 8.3), and the limitations (Sec 8.4).

8.1 The Roles of Users' Phone Use Purpose and Mental States in Smartphone Intervention

Most previous intervention techniques initiate interventions based on the amount of time and frequency of smartphone usage. However, quantifying smartphone use just by time oversimplifies and ignores the underlying causes. People use smartphones for work, study, and relaxation, as long as for meaningful reasons, the time is not a true problem. As Lukoff et al. found, even if participants didn't reduce their screen time, the intervention could make them feel better in the sense of agency and goal alignment, indicating users prioritize the quality of time over quantification [74]. *MindShift* initiates persuasion only when users recognize their current usage as habitual, aiming to enhance users' self-awareness of their habitual phone use behavior.

Additionally, certain mental states are linked to habitual smartphone use as a form of self-distraction. Although smartphone use serves as a coping mechanism for emotion fluctuations, studies show that habitually using them for escapism fails to effectively mitigate emotions [27]. Using smartphones for emotional regulation can lead to problematic smartphone use behavior, potentially leading to severe psychological issues such as depression [25, 133]. *MindShift* aims to intervene in habitual smartphone usage triggered by specific mental states. Our goal is to reduce users' problematic smartphone use and help individuals transition from avoidance-oriented coping to approach-oriented coping [24].

8.2 Towards Adaptive Persuasion Intervention

MindShift generates dynamic and personalized persuasion content by combining information such as users' simple physical contexts, mental states, and other behaviors. However, several participants still mentioned that the LLM-generated content sometimes could be "*stiff and templated*". This may be attributed to the limited prompt templates. Although the content generated by the LLM varies, the main theme is guided by our prompts, which could limit the variation of persuasion content. To improve, we suggest integrating user feedback into the system for more adaptive intervention. Currently, *MindShift* supports a simple thumb-up and thumb-down feedback mechanism. Even with such simple information, we could establish a human-in-the-loop setup to fine-tune content, aligning better with user preferences. Another aspect is to include more diverse behavior features captured by passive sensors on smartphones and wearables [40, 81, 127].

Moreover, future work can also consider collecting more comprehensive feedback from users. Users could customize the language style generated by an LLM, which can be coupled with adaptive

algorithms such as reinforcement learning to achieve a more intelligent just-in-time adaptive intervention (JITAI) system that evolves with users [32, 97].

8.3 Leveraging LLMs for Behavior Change

8.3.1 Advantages of Using LLMs for Behavior Change. Our study sheds light on the possibility of leveraging LLMs to change user behavior by influencing human cognition. Previous efforts in this field often hinge on users' ability to self-reflect and self-persuade, limited by the narrow scope of sentence databases used [128]. LLMs break this barrier, offering a broader range of persuasive strategies. They can generate adaptive and diverse persuasive content, tailored to the individual's context. In our study, we observed notable changes in cognition: participants' smartphone addiction scale scores dropped, and self-efficacy scores rose after the study (Figure 11a). This suggests a potential for long-term behavioral change, which we aim to explore further in our future work.

We also want to highlight that *MindShift* is just one example of the possibilities in this domain. Future research could integrate LLMs for more dynamic interventions, such as self-affirmation content generation in the typing intervention [128] and personalized visualizations [42]. Moreover, our methodology can potentially be expanded to other domains, such as smoking or alcohol cessation, eating diet, and physical activity promotion, where LLMs can be used to generate context-aware dynamic persuasive content for a specific well-being goal. We envision our work can inspire a number of creative LLM-powered intervention techniques in the future.

8.3.2 Design Implication for Using LLMs in Other Behavior Change Domains. Based on our findings in the study, we extract three design implications of using LLMs for behavior change in various domains.

First, investigating why people behave in certain ways is the basis of any intervention design, especially when LLMs can utilize such insights when generating persuasion. Our study explores the psychological factors behind habitual smartphone use. *MindShift* leveraging those factors outperformed *MindShift-Simple* only considering physical factors (see in Sec. 7.1.1 to 7.1.3).

Second, context is crucial for LLMs to generate dynamic, tailored persuasive messages. Our examples in Table 2 showcase the importance of user contexts for content generation. In our study, some user contexts, like mental states, cannot be detected automatically but depend on users' self-reports, which can be improved in future design. Past work has explored how to use smartphone usage data and machine learning to predict boredom and stress [21, 68, 81, 100, 115], and there have been researches using physiological measuring instruments to learn mental states from biosignals [112, 113, 124]. With the development of more smart and wearable technologies, there is the potential to track users' mental states automatically [31, 41]. This can simplify the intervention process and improve user experience. However, it remains an open question on how skipping self-reflection may impact the effectiveness of such a persuasion technique.

Last, crafting a suitable prompt is crucial for effectively incorporating expert knowledge into LLM generation. This often involves multiple attempts and adjustments to ensure the generated content aligns with expectations. While not the main contribution of our work, we conducted extensive iterations to ensure the appropriateness of the persuasive content. We have more discussion on the ethical concerns if prompt engineering is not done properly in the next paragraph. We refer future developers to recent studies, such as *EmotionPrompt* [69], for more comprehensive guidance on enhancing LLM outputs.

8.3.3 Ethical Concerns and Risk of Using LLMs for Behavior Change. Although *MindShift* performs well in changing problematic smartphone use, there are important ethical concerns we want to highlight about the risk of using LLMs for behavior change.

Despite carefully crafted prompts, developers face challenges ensuring the constant safety of generated persuasive messages. For example, while we fixed hallucination for our experiment, it is one of the biggest concerns in LLMs and can still possibly occur in real-world deployment [29, 47]. Additionally, although we didn't encounter it in our experiment, LLMs can generate dangerous content, such as abusive and discriminatory sentences. Furthermore, there is still room to improve LLMs' understanding of the nuances of human mental states [126]. For instance, if users of *MindShift* are already stressed due to their life objectives (*e.g.*, struggling with academic stress), additional reminders of these goals could exacerbate the stress or even cause harm. More future work is needed to improve safety and reduce the risk before we deploy LLMs for large-scale intervention studies.

Moreover, privacy is another critical concern since the detection of users' physical and psychological context data is needed. *MindShift* employs a commercial API from OpenAI, transmitting users' data to a third party. Although we intentionally designed the prompt to avoid including any identifiable information, there is still the risk of revealing information about their behavior and mental states. One solution for future study is to leverage open-sourced LLMs (such as LLaMA2 [117] or PaLM2 [6]) so that user's data can be appropriately handled and encrypted by ourselves instead of a third party.

8.4 Limitations

Our study has a few limitations. First, our experimental user group is limited to young adults, limiting result applicability. Future studies can expand the sample size and involve more diverse user groups. Second, our field experiment is short. A five-week deployment cannot reveal the longitudinal effect of such an intervention technique. Moreover, if our time and monetary budget allow, our experiment design can be improved by making the *Baseline* another two-week intervention session for a more fair comparison. Third, the validity and reliability test of mental state report questions needs further improvement. Testing convergent and discriminant validity, recruiting more samples for reliability tests, and using statistical methods to evaluate consistency are areas for enhancement. Additionally, our current method for detecting habitual use and mental states relies on self-reporting, increasing users' burden. We only consider initial habitual use upon users unlocking phones, neglecting shifts in user purposes during app usage. As we mentioned in Sec. 8.3.2, future work can explore automatic intent and mental state detection, and the data collected in this study can serve as a starting

point for machine learning training models. Finally, our study utilizes GPT-3.5 as a large language model, and its performance is still unstable. Future work can explore more lighted weighted and robust LLMs for local deployment, which can address the concerns mentioned in Sec. 8.3.3 to some extent.

9 CONCLUSION

This paper introduces *MindShift*, a mental-based persuasion intervention technique powered by LLM designed to mitigate problematic smartphone use. We conducted a Wizard-of-Oz study and an interview study to explore the mental states behind problematic smartphone use: *stress*, *boredom*, *inertia*, and designed four persuasion strategies: *understanding*, *comforting*, *evoking*, and *scaffolding habits*. *MindShift* (1) collects users' usage intent, usage behavior, physical context, mental states, goals&habits, (2) uses the persuasion strategies we design, (3) leverages LLMs to generate dynamic, personalized persuasion messages. Through a five-week within-subjects user experiment (N=25), we compared three intervention techniques (*MindShift*, *MindShift-Simple*, *Baseline*). *MindShift* outperforms *MindShift-Simple* and *Baseline*, improving acceptance rates (4.7-22.5%) and reducing app usage (7.4-9.8%). Notably, *MindShift* and *MindShift-Simple* significantly reduce SAS scores (34.7-25.8%) and increase self-efficacy scores (10.7-10.4%). Finally, users' subjective comments also confirm a perceived reduction in smartphone dependency and a willingness to continue to use *MindShift*. Our work provides valuable insights into the mental states behind problematic smartphone use and the effectiveness of LLMs-powered persuasion for smartphone intervention.

REFERENCES

[1] Dolores Albarracín, Aashna Sunderrajan, Sophie Lohmann, Man-Pui Sally Chan, and Duo Jiang. 2018. The psychology of attitudes, motivation, and persuasion. In *The handbook of attitudes, volume 1: Basic principles*. Routledge, 3–44.

[2] Carlos Alós-Ferrer, Sabine Hügelschäfer, and Jiahui Li. 2016. Inertia and decision making. *Frontiers in psychology* 7 (2016), 169.

[3] Amap. 2023. Amap API website. (2023). https://lbs.amap.com/.

[4] Ian A Anderson and Wendy Wood. 2021. Habits and the electronic herd: The psychology behind social media's successes and failures. *Consumer Psychology Review* 4, 1 (2021), 83–99.

[5] Ionut Andone, Konrad Blaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. Menthal: quantifying smartphone usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 559–564.

[6] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[7] Forest APP. 2021. Stay focused, be present. (2021). https://www.forestapp.cc/.

[8] Apple. 2021. About the security content of iOS 14.1 and iPadOS 14.1. (2021). https://support.apple.com/en-us/HT208982.

[9] Md Arefin, Md Islam, Mohitul Mustafi, Sharmina Afrin, Nazrul Islam, et al. 2018. Impact of smartphone addiction on academic performance of business students: A case study. *Md. and Mustafi, Mohitul and Afrin, Sharmina and Islam, Nazrul, Impact of Smartphone Addiction on Academic Performance of Business Students: A Case Study (August 21, 2018)* (2018).

[10] Cecil A Arnolds and Christo Boshoff. 2002. Compensation, esteem valence and job performance: an empirical assessment of Alderfer's ERG theory. *International Journal of Human Resource Management* 13, 4 (2002), 697–719.

[11] Amanda Baughan, Mingrui Ray Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa D Butler, and Alexis Hiniker. 2022. "I Don't Even Remember What I Read": How Design Influences Dissociation on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[12] Joseph B Bayer and Robert LaRose. 2018. Technology habits: Progress, problems, and prospects. *The psychology of habit: Theory, mechanisms, change, and contexts* (2018), 111–130.

[13] Stefan F Bernritter, Iris van Ooijen, and Barbara CN Müller. 2017. Self-persuasion as marketing technique: the role of consumers' involvement. *European Journal of Marketing* 51, 5/6 (2017), 1075–1090.

[14] Lauren G Block and Punam Anand Keller. 1997. Effects of self-efficacy and vividness on the persuasiveness of health communications. *Journal of consumer psychology* 6, 1 (1997), 31–54.

[15] Jane R Caulton. 2012. The development and use of the theory of ERG: A literature review. *Emerging Leadership Journeys* 5, 1 (2012), 2–8.

[16] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. From Gap to Synergy: Enhancing Contextual Understanding through Human-Machine Collaboration in Personalized Systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[17] Shao-I Chiu. 2014. The relationship between life stress and smartphone addiction on Taiwanese university student: A mediation model of learning self-efficacy and social self-efficacy. *Computers in human behavior* 34 (2014), 49–57.

[18] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing* 16, 1 (2017), 74–84.

[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[20] Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*. Vol. 55. Collins New York.

[21] Matteo Ciman and Katarzyna Wac. 2016. Individuals' stress assessment using human-smartphone interaction analysis. *IEEE Transactions on Affective Computing* 9, 1 (2016), 51–65.

[22] Russell B Clayton, Glenn Leshner, and Anthony Almond. 2015. The extended iSelf: The impact of iPhone separation on cognition, emotion, and physiology. *Journal of computer-mediated communication* 20, 2 (2015), 119–135.

[23] Emily IM Collins, Anna L Cox, Jon Bird, and Cassie Cornish-Tresstail. 2014. Barriers to engagement with a personal informatics productivity tool. In *Proceedings of the 26th Australian Computer-Human interaction Conference on Designing futures: The future of design*. 370–379.

[24] Bruce E Compas, Jennifer K Connor-Smith, Heidi Saltzman, Alexandria Harding Thomsen, and Martha E Wadsworth. 2001. Coping with stress during childhood and adolescence: problems, progress, and potential in theory and research. *Psychological bulletin* 127, 1 (2001), 87.

[25] Sarah M Coyne, Jane Shawcroft, Megan Gale, Douglas A Gentile, Jordan T Etherington, Hailey Holmgren, and Laura Stockdale. 2021. Tantrums, toddlers and technology: Temperament, media emotion regulation, and problematic media use in early childhood. *Computers in Human Behavior* 120 (2021), 106762.

[26] Élish Duke and Christian Montag. 2017. Smartphone addiction, daily interruptions and self-reported productivity. *Addictive behaviors reports* 6 (2017), 90–95.

[27] Megan Duvenage, Helen Correia, Bep Uink, Bonnie L Barber, Caroline L Donovan, and Kathryn L Modecki. 2020. Technology can sting when reality bites: Adolescents' frequent online coping is ineffective with momentary stress. *Computers in Human Behavior* 102 (2020), 248–259.

[28] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Xuhai Xu, Chun-Cheng Chang, Lianhui Qin, Xin Liu, Shwetak Patel, Vikram Iyer, et al. 2023. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *arXiv preprint arXiv:2311.13063* (2023).

[29] Xavier Ferrer, Tom van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. 2021. Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine* 40, 2 (2021), 72–80.

[30] Cynthia D Fisherl. 1993. Boredom at work: A neglected concept. *Human relations* 46, 3 (1993), 395–417.

[31] Shruti Gedam and Sanchita Paul. 2021. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access* 9 (2021), 84045–84066.

[32] Stephanie P Goldstein, Brittney C Evans, Daniel Flack, Adrienne Juarascio, Stephanie Manasse, Fengqing Zhang, and Evan M Forman. 2017. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. *International journal of behavioral medicine* 24 (2017), 673–682.

[33] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[34] Martin S Hagger. 2016. Non-conscious processes and dual-process theories in health psychology. *Health Psychology Review* 10, 4 (2016), 375–380.

[35] Jaap Ham and Sitwat Usman Langrial. 2020. Learning to Stop Smoking: Understanding Persuasive Applications' Long-Term Behavior Change Effectiveness Through User Achievement Motivation. In *Persuasive Technology. Designing*

for Future Change: 15th International Conference on Persuasive Technology, PERSUASIVE 2020, Aalborg, Denmark, April 20–23, 2020, Proceedings 15. Springer, 139–149.

[36] Daniel Harrison, Paul Marshall, Nadia Bianchi-Berthouze, and Jon Bird. 2015. Activity tracking: barriers, workarounds and customisation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 617–621.

[37] Andree Hartanto and Hwajin Yang. 2016. Is the smartphone a smart choice? The effect of smartphone separation on executive functions. *Computers in human behavior* 64 (2016), 329–336.

[38] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509* (2022).

[39] Joshua Harwood, Julian J Dooley, Adrian J Scott, and Richard Joiner. 2014. Constantly connected–The effects of smart-devices on mental health. *Computers in Human Behavior* 34 (2014), 267–272.

[40] Liang He, Ruolin Wang, and Xuhai Xu. 2020. PneuFetch: supporting blind and visually impaired people to fetch nearby objects via light haptic cues. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[41] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors* 21, 10 (2021), 3461.

[42] Alexis Hiniker, Sungsoo Hong, Tadayoshi Kohno, and Julie A Kientz. 2016. MyTime: designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4746–4757.

[43] Alexis Hiniker, Shwetak N Patel, Tadayoshi Kohno, and Julie A Kientz. 2016. Why would you do that? predicting the uses and gratifications behind smartphone-usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 634–645.

[44] Wilhelm Hofmann, Malte Friese, and Fritz Strack. 2009. Impulse and self-control from a dual-systems perspective. *Perspectives on psychological science* 4, 2 (2009), 162–176.

[45] Kyung-Hye Hwang, Yang-Sook Yoo, and Ok-Hee Cho. 2012. Smartphone overuse and upper extremity pain, anxiety, depression, and interpersonal relationships among college students. *The Journal of the Korea Contents Association* 12, 10 (2012), 365–375.

[46] Bytedance Inc. 2023. Tiktok App. (2023). https://www.tiktok.com/.

[47] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[48] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[50] Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. Adaptive persuasive systems: a study of tailored persuasive text messages to reduce snacking. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 2 (2012), 1–25.

[51] Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies* 77 (2015), 38–51.

[52] Alan E Kazdin. 1982. Observer effects: Reactivity of direct observation. *New Directions for Methodology of Social & Behavioral Science* (1982).

[53] Inyeop Kim, Gyuwon Jung, Hayoung Jung, Minsam Ko, and Uichin Lee. 2017. Let's focus: location-based intervention tool to mitigate phone use in college classrooms. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 101–104.

[54] Jaejeung Kim, Hayoung Jung, Minsam Ko, and Uichin Lee. 2019. Goalkeeper: Exploring interaction lockout mechanisms for regulating smartphone use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–29.

[55] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. 2019. LocknType: Lockout task intervention for discouraging smartphone app use. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[56] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. *arXiv preprint arXiv:2401.06866* (2024).

[57] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 272–283.

[58] Minsam Ko, Subin Yang, Joonwon Lee, Christian Heizmann, Jinyoung Jeong, Uichin Lee, Daehee Shin, Koji Yatani, Junehwa Song, and Kyong-Mee Chung. 2015. NUGU: a group-based intervention app for improving self-regulation of limiting smartphone use. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1235–1245.

[59] Jaap M Koolhaas, Alessandro Bartolomucci, Bauke Buwalda, Seitse F de Boer, Gabriele Flügge, S Mechiel Korte, Peter Meerlo, Robert Murison, Berend Olivier, Paola Palanza, et al. 2011. Stress revisited: a critical evaluation of the stress concept. *Neuroscience & Biobehavioral Reviews* 35, 5 (2011), 1291–1301.

[60] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[61] Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman A. S. Farb, and Joseph Jay Williams. 2023. Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 129, 7 pages. https://doi.org/10.1145/3544549.3585614

[62] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[63] Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science* 21, 7 (2010), 984–991.

[64] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. 2013. Development and validation of a smartphone addiction scale (SAS). *PloS one* 8, 2 (2013), e56936.

[65] Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727* (2023).

[66] Liette Lapointe, Camille Boudreau-Pinsonneault, and Isaac Vaghefi. 2013. Is smartphone usage truly smart? A qualitative investigation of IT addictive behaviors. In *2013 46th Hawaii international conference on system sciences*. IEEE, 1063–1072.

[67] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. 2014. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2327–2336.

[68] Damien Lekkas, George D Price, and Nicholas C Jacobson. 2022. Using smartphone app use and lagged-ensemble machine learning for the prediction of work fatigue and boredom. *Computers in human behavior* 127 (2022), 107029.

[69] Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760* (2023).

[70] Yu-Hsuan Lin, Li-Ren Chang, Yang-Han Lee, Hsien-Wei Tseng, Terry BJ Kuo, and Sue-Huei Chen. 2014. Development and validation of the Smartphone Addiction Inventory (SPAI). *PloS one* 9, 6 (2014), e98312.

[71] Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636.

[72] Markus Löchtefeld, Matthias Böhmer, and Lyubomir Ganev. 2013. AppDetox: helping users with mobile app addiction. In *Proceedings of the 12th international conference on mobile and ubiquitous multimedia*. 1–2.

[73] Tao Lu, Hongxiao Zheng, Tianying Zhang, Xuhai Xu, and Anhong Guo. 2024. InteractOut: Leveraging Interaction Proxies as Input Manipulation Strategies for Reducing Smartphone Overuse. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[74] Kai Lukoff, Ulrik Lyngs, Karina Shirokova, Raveena Rao, Larry Tian, Himanshu Zade, Sean A. Munson, and Alexis Hiniker. 2023. SwitchTube: A Proof-of-Concept System Introducing "Adaptable Commitment Interfaces" as a Tool for Digital Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 197, 22 pages. https://doi.org/10.1145/3544548.3580703

[75] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. 2021. How the design of youtube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[76] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What makes smartphone use meaningful or meaningless? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[77] Ulrik Lyngs, Kai Lukoff, Laura Csuka, Petr Slovák, Max Van Kleek, and Nigel Shadbolt. 2022. The Goldilocks level of support: Using user reviews, ratings, and installation numbers to investigate digital self-control tools. *International*

journal of human-computer studies 166 (2022), 102869.
[78] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-control in cyberspace: Applying dual systems theory to a review of digital self-control tools. In *proceedings of the 2019 CHI conference on human factors in computing systems*. 1–18.
[79] Ulrik Lyngs, Kai Lukoff, Petr Slovak, William Seymour, Helena Webb, Marina Jirotka, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2020. 'I Just Want to Hack Myself to Not Get Distracted' Evaluating Design Interventions for Self-Control on Facebook. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[80] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).
[81] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–32.
[82] William L Mikulas and Stephen J Vodanovich. 1993. The essence of boredom. *The Psychological Record* 43, 1 (1993), 3.
[83] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* (2021).
[84] Lewis Mitchell and Zaheer Hussain. 2018. Predictors of problematic smartphone use: An examination of the integrative pathways model and the role of age, gender, impulsiveness, excessive reassurance seeking, extraversion, and depression. *Behavioral Sciences* 8, 8 (2018), 74.
[85] Alberto Monge Roffarello and Luigi De Russis. 2019. The race towards digital wellbeing: Issues and opportunities. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[86] Alberto Monge Roffarello and Luigi De Russis. 2023. Nudging Users or Redesigning Interfaces? Evaluating Novel Strategies for Digital Wellbeing Through inControl. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 100–109.
[87] Alberto Monge Roffarello, Kai Lukoff, Luigi De Russis, et al. 2023. Defining and Identifying Attention Capture Damaging Patterns in Digital Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 1–30.
[88] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 205–212.
[89] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6 (2018), e10148.
[90] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2018), 446–462.
[91] Ulrike E Nett, Thomas Goetz, and Lia M Daniels. 2010. What to do when feeling bored?: Students' strategies for coping with boredom. *Learning and Individual Differences* 20, 6 (2010), 626–638.
[92] Chukwuemeka Nwagu. 2023. Design and Evaluation of the Chai Wallpaper: A Mindfulness-Based Persuasive Intervention for Absent-Minded Smartphone Use. (2023).
[93] Fabian Okeke, Michael Sobolev, Nicola Dell, and Deborah Estrin. 2018. Good vibrations: can a digital nudge reduce digital overload?. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*. 1–12.
[94] OpenAI. 2022. Introducing ChatGPT. (2022). https://openai.com/blog/chatgpt.
[95] OpenAI. 2023. GPT best practices. (2023). https://platform.openai.com/docs/guides/gpt-best-practices.
[96] Rita Orji and Karyn Moffatt. 2018. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health informatics journal* 24, 1 (2018), 66–91.
[97] Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K. Dey, and Xuhai Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642747
[98] Hancheol Park and Gahgene Gweon. 2015. Initiating moderation in problematic smartphone usage patterns. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1585–1590.
[99] Joonyoung Park, Jin Yong Sim, Jaejeung Kim, Mun Yong Yi, and Uichin Lee. 2018. Interaction restraint: enforcing adaptive cognitive tasks to restrain problematic user interaction. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[100] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 825–836.
[101] Charlie Pinder, Jo Vermeulen, Benjamin R Cowan, and Russell Beale. 2018. Digital behaviour change interventions to break and form habits. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 3 (2018), 1–66.
[102] Aarathi Prasad, Lucas S LaFreniere, Vaasu Taneja, and Zoe Beals. 2021. Addressing Problematic Smartphone Use with a Personalized, Goal-based Approach. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 131–134.
[103] Paul C Price, Rajiv S Jhangiani, and I-Chant A Chiang. 2015. Reliability and validity of measurement. *Research methods in psychology* (2015).
[104] Aditya Kumar Purohit, Louis Barclay, and Adrian Holzer. 2020. Designing for digital detox: Making social media less addictive with digital nudges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.
[105] Aditya Kumar Purohit, Torben Jan Barev, Sofia Schöbel, Andreas Janson, and Adrian Holzer. 2023. Designing for DigitalWellbeing on a Smartphone: Co-creation of Digital Nudges to Mitigate Instagram Overuse. (2023).
[106] Aditya Kumar Purohit, Kristoffer Bergram, Louis Barclay, Valéry Bezençon, and Adrian Holzer. 2023. Starving the Newsfeed for Social Media Detox: Effects of Strict and Self-regulated Facebook Newsfeed Diets. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[107] Aditya Kumar Purohit and Adrian Holzer. 2019. Functional digital nudges: Identifying optimal timing for effective behavior change. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
[108] Christof Rapp. 2002. Aristotle's rhetoric. (2002).
[109] Alan M Rubin. 2009. Uses and gratifications. *The SAGE handbook of media processes and effects* (2009), 147–159.
[110] Maya Samaha and Nazir S Hawi. 2016. Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in human behavior* 57 (2016), 321–325.
[111] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized self-efficacy scale. *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs* 35 (1995), 37.
[112] Jungryul Seo, Teemu H Laine, and Kyung-Ah Sohn. 2019. An exploration of machine learning methods for robust boredom classification using EEG and GSR data. *Sensors* 19, 20 (2019), 4561.
[113] Jungryul Seo, Teemu H Laine, and Kyung-Ah Sohn. 2019. Machine learning approaches for boredom classification using EEG. *Journal of Ambient Intelligence and Humanized Computing* 10 (2019), 3831–3846.
[114] Herbert W Simons. 1976. Persuasion: Understanding, practice, and analysis. *(No Title)* (1976).
[115] Thomas Stütz, Thomas Kowar, Michael Kager, Martin Tiefengrabner, Markus Stuppner, Jens Blechert, Frank H Wilhelm, and Simon Ginzinger. 2015. Smartphone based stress prediction. In *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29–July 3, 2015. Proceedings 23*. Springer, 240–251.
[116] Kelly J Thomas Craig, Laura C Morgan, Ching-Hua Chen, Susan Michie, Nicole Fusco, Jane L Snowdon, Elisabeth Scheufele, Thomas Gagliardi, and Stewart Sill. 2021. Systematic review of context-aware digital behavior change interventions to improve health. *Translational behavioral medicine* 11, 5 (2021), 1037–1048.
[117] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[118] Jonathan A Tran, Katie S Yang, Katie Davis, and Alexis Hiniker. 2019. Modeling the engagement-disengagement cycle of compulsive phone use. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[119] Zahra Vahedi and Alyssa Saiphoo. 2018. The association between smartphone use, stress, and anxiety: A meta-analytic review. *Stress and Health* 34, 3 (2018), 347–358.
[120] Chuang Wang, Matthew Lee, and Zhongsheng Hua. 2014. Understanding and predicting compulsive smartphone use: An extension of reinforcement sensitivity approach. (2014).
[121] Chuang Wang and Matthew KO Lee. 2020. Why we cannot resist our smartphones: investigating compulsive use of mobile SNS from a Stimulus-Response-Reinforcement perspective. *Journal of the Association for Information Systems* 21, 1 (2020), 4.
[122] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported

Data.

[123] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time' Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1729–1738.

[124] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2019. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–33.

[125] Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2021. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[126] Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. arXiv:2307.14385 [cs.HC]

[127] 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 E Morris, Eve Riskin, Jennifer Mankoff, and Anind K Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 18.

[128] Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, Ruolin Wang, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, and Anind K Dey. 2022. TypeOut: Leveraging Just-in-Time Self-Affirmation for Smartphone Overuse Reduction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[129] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Shao Zhang, Ethan Rogers, Stephen Intille, Nawar Shara, Dakuo Wang, et al. 2023. Talk2Care: Facilitating Asynchronous Patient-Provider Communication with Large-Language-Model. *arXiv preprint arXiv:2309.09357* (2023).

[130] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).

[131] Mingrui Ray Zhang, Kai Lukoff, Raveena Rao, Amanda Baughan, and Alexis Hiniker. 2022. Monitoring Screen Time or Redesigning It? Two Approaches to Supporting Intentional Social Media Use. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[132] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).

[133] Andras N Zsido, Nikolett Arato, Andras Lang, Beatrix Labadi, Diana Stecina, and Szabolcs A Bandi. 2021. The role of maladaptive cognitive emotion regulation strategies and social anxiety in problematic smartphone and social media use. *Personality and Individual Differences* 173 (2021), 110647.

APPENDIX

Table 4: Persuasive Messages in WoZ Study. Four types of persuasive messages delivered in WoZ study and their examples.

Types	Examples
Usage Notice	*"You have used Wechat for 2 hours and 25 minutes today. Put down your phone please!"*
Practical Guidance	*"Are you still using WeChat? Have you completed the task of analyzing data today?"*
Encouragement	*"You have only spent 2 hours on your phone today, that's excellent! Keep up the good work~"*
Deterrent	*"Using the phone before bedtime can affect the quality of your sleep."*

Table 5: Takeaways from WoZ & semi-structured interview studies. (E) shows messages sent by experimenters, (W) demonstrates participants' reflection quotes in the WoZ study, and (S) means participants' quotes in the semi-structured interview study.

WoZ study	
Types of smartphone use	**Representative quote(s)**
Instrumental use (not to be intervened)	*"You've already spent one and a half an hour on WeChat today. Please put your phone down and focus on other aspects of life. (E)"* *"I felt a bit resentful because I was using WeChat to manage my affairs, rather than idly wasting time. (W1)"*
Instrumental use - relaxation (not to be intervened)	*"Please stop browsing Zhihu and engage in more meaningful activities. (E)"* *"I don't agree. Finding joy in browsing Zhihu constitutes meaning for me. (W5)"*
Habitual use (to be intervened)	*"You have used Zhihu for 2 hours today. Think about what else you have to do tonight. (E)"* *"Thanks for this suggestion. I always failed to control myself to open Zhihu. (W10)"*
Factors affecting persuasion effectiveness	**Representative quote(s)**
Mental states	*"When you find yourself with idle time, consider engaging in meaningful activities such as reading, writing, or drawing. (E)"* *"It correctly identified my state of not knowing what to do and offered sensible advice. (W11)"*
Personal goals	*"Your WeChat session has lasted 10 minutes. Please set aside your device to alleviate eye strain. (E)"* *"Keeping the eyes healthy is one thing I really care about, so I like this advice. (W12)"*
Contextual information	*"The afternoon is a good time for studying. Don't spend too much time on your phone. (E)"* *"Afternoon is indeed my study time during which I should improve my efficiency. It is right. (W4)"*
Semi-structured interview study	
Mental states related to habitual use	**Representative quote(s)**
Boredom	*"I find myself instinctively reaching for my phone in search of mental stimulation when doing simple assignments light on cognitive engagement." (S1)*
Stress	*"One day, work wasn't progressing well and I was so frustrated that I unlocked my phone for a quick view to ease my mood." (S9)*
Inertia	*"Upon returning home after work, I intended to transition back into a focused state for reading or other activities but struggled to shift from a relaxed state. At that point, my phone was the tool for procrastination." (S2)*
Activity engagement states	**Representative quote(s)**
Engaging in activities	*"I was reluctant to start handling this challenging work that I scrolled my phone screen anxiously." (S2)*
Not engaging in activities	*"After getting off work and returning home, I collapse on the sofa and binge-watch Tiktok for one to two hours." (S9)*

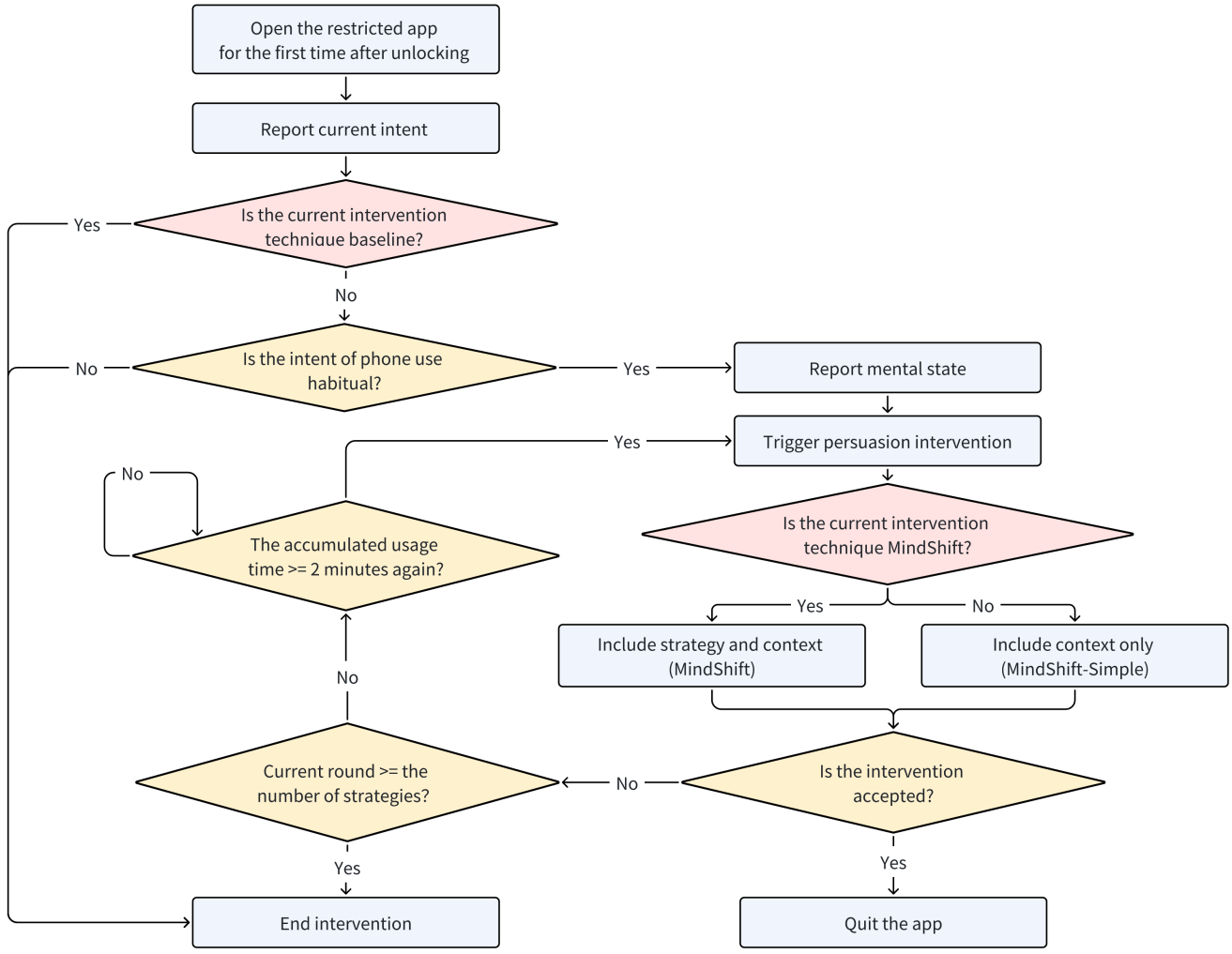


Figure 12: Interaction Process of Three Intervention Techniques. The two red process blocks illustrate the differences between the three intervention techniques.