Bloom: Designing for LLM-Augmented Behavior Change Interactions

Matthew Jörke
joerke@stanford.edu
Computer Science
Stanford University
Stanford, USA

Defne Genç*
defneg@stanford.edu
Computer Science
Stanford University
Stanford, USA

Valentin Teutschbein*
valentin.teutschbein@student.hpi.de
Computer Science
Hasso Plattner Institute
Potsdam, Germany

Shardul Sapkota
sapkota@stanford.edu
Computer Science
Stanford University
Stanford, USA

Sarah Chung
chvng@stanford.edu
Computer Science
Stanford University
Stanford, USA

Paul Schmiedmayer
schmiedmayer@stanford.edu
Mussallem Center for Biodesign
Stanford University
Stanford, USA

Maria Ines Campero
icampero@stanford.edu
School of Medicine
Stanford University
Stanford, USA

Abby C. King
king@stanford.edu
School of Medicine
Stanford University
Stanford, USA

Emma Brunskill
ebrun@cs.stanford.edu
Computer Science
Stanford University
Stanford, USA

James A. Landay
landay@stanford.edu
Computer Science
Stanford University
Stanford, USA

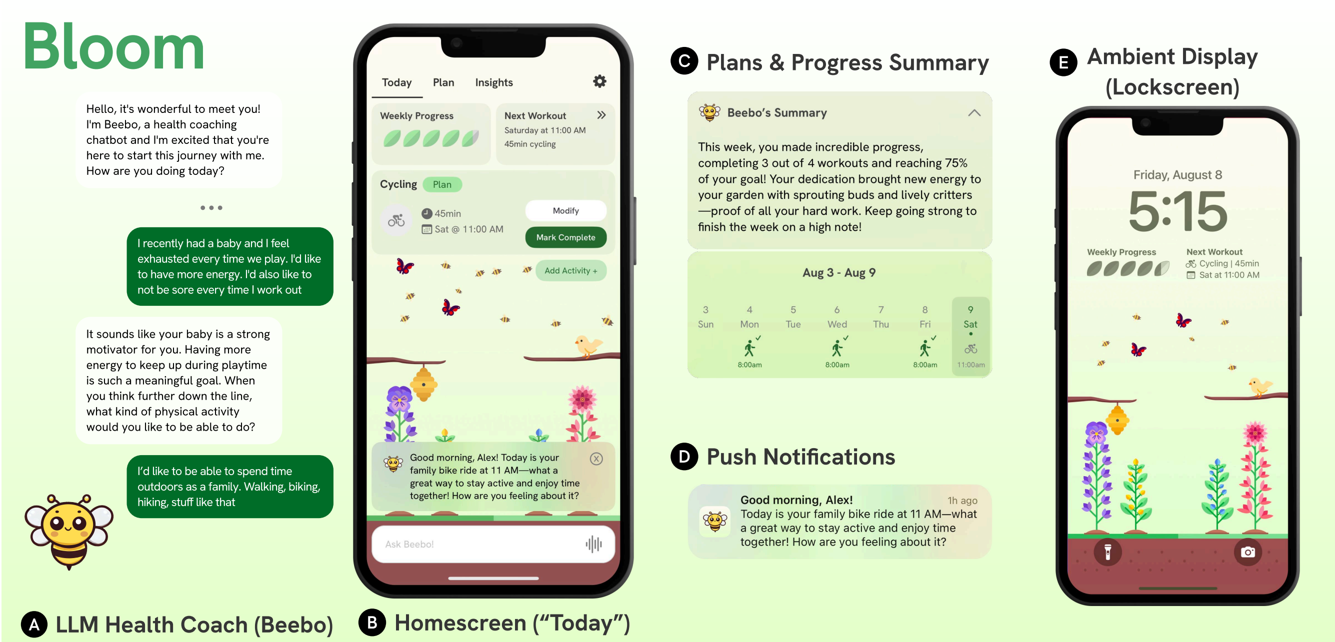

Figure 1: Bloom integrates an LLM-based physical activity coaching chatbot with established behavior change interactions. (A) The user begins with an onboarding conversation with Beebo, an LLM coaching chatbot that implements conversational strategies from motivational interviewing and a validated health coaching program. (B) The Today screen, the app's home screen, shows weekly progress, upcoming activities, and proactive messages from Beebo. Conversations with the agent elicit *qualitative* context, allowing Bloom to personalize various LLM-augmented behavior change interactions, such as (C) LLM-generated physical activity plans, accompanied by progress summaries, (D) LLM-generated notifications that encourage action or reflection, and (E) a garden-based ambient display that updates on the user's lockscreen.

Abstract

Large language models (LLMs) offer novel opportunities to support health behavior change, yet existing work has narrowly focused on text-only interactions. Building on decades of HCI research demonstrating the effectiveness of UI-based interactions, we present Bloom, an application for physical activity promotion that integrates an LLM-based health coaching chatbot with established UI-based interactions. As part of Bloom's development, we conducted a redteaming evaluation and contribute a safety benchmark dataset. In a four-week randomized field study (N=54) comparing Bloom to a non-LLM control, we observed important shifts in psychological outcomes: participants in the LLM condition reported stronger beliefs that activity was beneficial, greater enjoyment, and more self-compassion. Both conditions significantly increased physical activity levels, doubling the proportion of participants meeting recommended weekly guidelines, though we observed no significant differences between conditions. Instead, our findings suggest that LLMs may be more effective at shifting mindsets that precede longer-term behavior change.

1 Introduction

Regular physical activity (PA) reduces the risk of cardiovascular disease, diabetes, premature mortality, depression, and numerous other important physical and mental health outcomes [107], yet one in four adults worldwide and nearly half of the US population fall short of recommended PA guidelines [21, 41]. Many of the most effective approaches for increasing PA, such as in-person coaching, are resource-intensive and costly to deliver at scale. This challenge has motivated research in mobile health, which explores how more accessible technologies like smartphones and wearables can be used to deliver scalable, cost-effective, and personalized support [49].

Recent advances in large language models (LLMs) [17, 100] have sparked a growing interest in their applications for health behavior change, presenting novel opportunities for improved support. Early research on LLM-based health coaching chatbots suggest that they offer marked improvements in conversational flexibility over prior rule-based systems [48, 55, 132, 138]. This flexibility could allow LLMs to more accurately implement evidence-based health communication frameworks like motivational interviewing [86]. Moreover, LLMs offer an improved capacity to interpret diverse sources of personal context, enabling interventions that are more effectively tailored to an individual [57, 81].

Unlike prior systems, which primarily rely on quantitative context (e.g., step count) for personalization, LLMs can also leverage *qualitative context* [55] that is more easily expressed in natural language. Qualitative factors such as goals, motivation, values, life circumstances, time constraints, and access to resources are key constructs in behavior change theory and are crucial for effective personalization.

Meanwhile, decades of research in human-computer interaction (HCI) has produced a rich body of interactions for health behavior change applications—from goal-setting [36] and self-tracking [38] to ambient feedback [32, 91], and just-in-time nudges [94]—which are both grounded in behavior change theory [31, 75, 85] and have demonstrated efficacy [133, 139]. However, existing research on LLMs for behavioral health has focused almost exclusively on text-only interactions, either through chat [55, 132, 138], summaries [57, 81], or nudging [77, 119]. This narrow focus overlooks the potential for LLMs to *augment* established UI-based interactions, rather than replace them. Behavior change theory emphasizes that behavior is shaped not just by what is said, but when, how, and through which channels it is delivered [85, 88]. HCI research has shown that multimodal interaction [101, 127], which blends textual, graphical, ambient, or wearable delivered feedback, can help users manage cognitive load [102], make behaviors more salient through glanceable or ambient displays [32, 46, 79], and foster affective engagement [73, 93]. Taken together, this research suggests that behavior change interactions that augment LLM chat with other interaction modalities may offer greater efficacy.

At the same time, the technology is recent and real-world evaluations of LLMs for behavioral health remain limited. Most studies have relied on static assessments or single-session studies [48, 55, 77, 82, 120], while the few field trials conducted have included fewer than 20 participants [132, 138]. Importantly, most prior work lacks rigorous controls and assumes the efficacy of LLMs without directly evaluating LLM-based systems relative to established pre-LLM approaches. As a result, existing findings provide limited insights into the design requirements for LLM-augmented interactions or longitudinal LLM coaching.

To address this gap, we present **Bloom**, an application for promoting PA that combines an LLM coaching chatbot with established UI-based behavior change interactions, including goal-setting, action planning, activity tracking, data visualization, an ambient display, and push notifications (Figure 1). Our goal in developing and evaluating Bloom was to surface design insights for *LLM-augmented behavior change interactions* and establish insights into their efficacy. Bloom's LLM coach implements conversational strategies from motivational interviewing [86] and implements a scientifically validated health coaching program [60]. Conversations with Bloom's chatbot thus simultaneously serve as standalone interventions and as sources of qualitative context for other parts of the system to leverage for personalization. As part of Bloom's development, we also conducted a redteaming evaluation with four domain experts and created a safety benchmark dataset for LLM coaching with 600 examples to evaluate our system's safety filters.

We evaluate Bloom in a four-week, between-subjects, randomized field study with $N = 54$ participants, comparing Bloom to a no-LLM control that removes the LLM coach and all LLM augmentation. To the best of our knowledge, this is the first field study evaluation of LLM-based PA coaching. Our comparative study was formative and design-oriented, with the no-LLM control condition allowing us to isolate changes in behavior and experiences specifically attributable to LLM augmentation, as well as whether, when, and how LLMs provide benefits beyond well-designed, evidence-based, pre-LLM systems. Our evaluation uses a mixed-methods approach, including qualitative coding of semi-structured interviews; pre/post, weekly, and daily survey measures; objective PA data collected from participants' wearables; as well as application usage logs including chats, plans, and UI interactions.

We observed important shifts in psychological and motivational outcomes, with participants in the LLM condition reporting stronger beliefs that their activity was beneficial to their health. They also

*Both authors contributed equally to this research.

noted greater enjoyment of exercise, an expanded appreciation of what "counts" as activity, and increased self-compassion when goals were missed. The LLM condition also produced more varied and personalized plans, with slightly higher completion rates, and participants spent more than five times as much time in the app than control participants. Both conditions significantly increased their PA relative to a pre-study baseline, doubling the number of participants who met or exceeded the recommended 150 min/week of PA. However, we found no statistically significant differences in quantitative wearable PA outcomes between conditions during our four-week study period. Instead, our survey and interview findings suggest that LLMs may be particularly valuable for fostering positive mindsets, flexible planning, and sustained engagement—factors that behavior change theory links to longer-term maintenance rather than as immediate drivers of short-term levels of PA. We discuss the implications of these findings for designing LLM-augmented interventions and for understanding the role of LLMs in behavioral health systems.

In summary, this work makes the following contributions:

(1) The **Bloom system**, which implements a novel design for integrating LLM chat with established, UI-based behavior change interactions. Bloom is grounded in evidence-based strategies from behavior change theory and interactions from HCI research. Bloom leverages qualitative context from coaching conversations to personalize a number of *LLM-augmented behavior change interactions*.
(2) The **first safety evaluation of an LLM-based health coaching agent**. We share results, prompts, and a benchmark dataset containing 600 examples to support future research on safety.
(3) Findings from the **first field study of an LLM-based PA coaching agent**, a four-week, between-subjects randomized study with $N = 54$ participants comparing Bloom to a non-LLM control. We analyze wearable PA outcomes, survey responses, system interaction logs, and qualitative coding of pre/post semi-structured interviews to assess both behavioral and psychological outcomes.
(4) **Design implications** drawn from our field study's findings to inform future work on LLM-augmented behavior change interactions and LLM health coaching systems.

2 Related Work

In this section, we situate our work within the prior literature on HCI systems for PA behavior change, automated and human health coaching, and LLMs for behavioral health.

2.1 HCI Systems for Physical Activity Promotion

Designing for PA promotion has a long history in HCI [38], with many approaches advocated in the literature [30, 49, 64]. One popular line of work draws from the broader field of personal informatics [71] and aims to promote behavior change by helping users make sense of their activity patterns through self-monitoring [25, 65] and reflection [7, 12, 24]. Common implementations frequently include visualizations [4, 26, 56, 125] such as step count graphs, often seen in commercial health-tracking apps. Other systems employ more qualitative forms of feedback, including ambient displays [32, 70, 91] and textual feedback [10, 29]. Qualitative feedback can be less overwhelming [35] and promote more positive mindsets [93] than quantitative feedback. Nudging represents another class of interventions, which in mobile health are often instantiated as push notification reminders for just-in-time support [9, 52, 94, 128]. Unlike other kinds of reflective or ambient support, just-in-time interventions aim to capture the user's attention and often optimize intervention timing and content to maximize efficacy.

Another prominent body of work focuses on goal-setting as a strategy for promoting PA [36, 90]. Goal-setting theory emphasizes that goals are most effective when they are personally meaningful, realistic, and sufficiently challenging [75]. Users are often asked to set their own PA goals [42, 73], or are alternatively provided a goal based on their behavior history [29]. Prior work has also employed conversational agents for collaborative goal-setting [14, 55, 58]. Since intentions alone do not always translate to goal achievement, action plans [1] and implementation intentions [44, 45, 114] are popular techniques for facilitating goal realization by specifying when, where, and how a behavior will occur.

Beyond individual-level interventions, prior work has also explored social strategies for behavior change. These strategies typically involve sharing within a social network [3, 29, 73, 126] or incorporating leaderboards and challenges to introduce competition in games [2, 20, 43, 116, 117]. While collaborative approaches have been shown to improve engagement and activity levels in some contexts, social comparison can sometimes feel demotivating or misaligned with preferences [90, 92], making social features challenging to design effectively. Competitive elements in particular, such as leaderboards or challenges, have shown mixed effects on users' motivation, especially during periods when they were less active [29, 62]. Although gamification can increase short-term engagement, its long-term impact on sustained behavior change also remains underexplored [122].

Bloom implements several interactions from this literature, including activity tracking, data visualizations, ambient displays, push notifications, goal setting, and action planning. Due to the challenge of appropriately designing for social and gamification interactions, compounded with the uncertainty of designing for novel, LLM-augmented interactions, we exclude these interactions, leaving them as important explorations for future work.

2.2 Health Coaching

Health coaching is an established intervention for facilitating behavior change [98, 99, 136]. Although coaching is a multi-disciplinary concept with numerous everyday associations (e.g., sports, executive, or life coaching), here we adopt Olsen et al.'s [98] definition of health coaching as *"a goal-oriented, client-centered partnership that is health-focused and occurs through a process of client enlightenment and empowerment."* A common framework used in such programs is motivational interviewing (MI) [86], an evidence-based counseling approach that emphasizes eliciting motivation from the client rather than imposing prescriptive advice.

However, in-person health coaching is expensive, not widely accessible, and does not scale to global need. A long history of research on health dialogue systems [13] and automated health coaching [87] has aimed to address these challenges by emulating aspects of human coaching with conversational agents. This research dates back at least as far as the seminal work on the Transtheoretical Model (TTM) itself [108], and we refer to [55, 87] for reviews of pre-LLM automated coaching. While meta-analyses have found that pre-LLM coaching chatbots are effective at promoting physical activity [118], these systems rely on predefined rules or templates that inherently limit their flexibility and personalization compared to human coaches [76]. Unlike prior rule-based systems, Bloom leverages an LLM to engage in dynamic, open-ended conversations, interpret and integrate user data through tool calls, and reference past interactions.

2.3 Large Language Models and Behavioral Health

Given LLMs offer marked improvements over prior rule-based chatbots in conversational flexibility, recent work has explored their applications to health coaching. Bloom's LLM health coach is based on the GPTCoach system [55], a GPT-4-based chatbot that implements the onboarding conversation of a physical activity coaching program and uses conversational strategies from MI. Participants reported highly positive and personalized experiences interacting with GPTCoach, and the authors found that it was effective at using strategies aligned with MI principles. Other LLM-based chatbots have explored domains beyond physical activity. Wang et al. [132] introduced HealthGuru, a sleep health chatbot that uses behavioral guidelines and sensor data to provide recommendations. Xu et al. [138] developed a chatbot for New Year's resolutions that personalized suggestions based on chatbot interaction, user profile, and contextual cues (e.g., location, time, goals, etc.). In concurrent work, Heydari et al. [48] present a multi-agent system consisting of a data science agent, health domain expert agent, and a health coaching agent. Bloom extends these approaches by integrating chat with a suite of UI-based behavior change interactions.

A parallel line of research has focused on improving LLMs' ability to engage in MI. Studies have applied MI principles in health coaching contexts [55, 82] and beyond, including mental health [23, 115], substance use counseling [68, 121], and wellness-related reflection [8, 111]. More recent work has enhanced MI adherence through strategy prediction [124] or few-shot demonstrations [137]. Bloom's LLM coach conditions each response on a predicted MI strategy.

Beyond chatbots, most studies on LLMs for behavioral health have focused on improving message personalization or sensor data analysis. For example, to improve adherence to the Capacity, Opportunity, Motivation-Behavior (COM-B) framework, Vardhan et al. [129] explored LLM prompting techniques, whereas Mantena et al. [77] used a fine-tuning approach to align with the TTM. LLMs have also been applied to downstream health tasks, including generating stage-appropriate behavior change advice [5], modeling trust and self-referencing in coaching [83], performing health tasks on raw self-tracking data [27, 37, 72, 74], and extracting insights from personal health data [39, 40, 51, 57, 81, 123, 142, 143]. While these systems show improvement in adhering to theoretical frameworks and personalizing content with quantitative data, they are not designed to proactively elicit and incorporate qualitative information about users' goals, values, or barriers into a longitudinal coaching process. Bloom explicitly integrates multi-session LLM coaching with behavior change interventions, leveraging the qualitative context elicited from coaching for greater personalization.

3 Bloom: System Overview

Bloom is an iOS application designed for multi-week PA support that integrates an LLM coaching agent (Beebo) with UI-based behavior change interactions. In this section, we describe the Bloom system, including the Active Choices program and Bloom's behavior change interactions, system architecture, LLM coaching agent, and safety filters. Bloom's user interface is shown in Figures 1 and 2 and our system architecture is shown in Figure 3. Our code is publicly available at https://github.com/StanfordHCI/Bloom.

3.1 Active Choices Program

Bloom implements the Stanford Active Choices Program [60], an evidence-based, scientifically validated counseling program for PA promotion developed by researchers at the Stanford University School of Medicine [22, 59, 61, 134, 135]. Active Choices is grounded in the Transtheoretical Model [108] and Social Cognitive Theory [6], and many Active Choices facilitators are trained in MI [86]. Bloom's LLM coaching agent directly draws on the program's structure for onboarding conversations, exercise planning, and check-in conversations. To adapt Active Choices to a digital format, Bloom adds several additional behavior change interactions, described in the following section.

3.2 Behavior Change Interactions

Bloom integrates the LLM coaching agent with six evidence-based behavior change interactions, leveraging the qualitative context elicited from the agent for richer personalization. In Table 1, we list the behavior change techniques (BCTs), drawn from Michie et al.'s taxonomy [84], implemented by each interaction. Whereas interactions are described at the level of the user interface, BCTs capture the underlying "active components" of behavior change interventions and help ground our interactions in techniques known to be effective from behavior change theory.

3.2.1 Goal-Setting & Planning. Beebo, Bloom's LLM agent, guides the user through weekly, collaborative goal-setting. A weekly goal is an exercise plan that specifies FITT (frequency, intensity, time, type) parameters for each activity (e.g., a moderate-intensity, 20-min walk at 8AM on Mon-Wed-Fri). Beebo first inquires about motivations, long-term goals, past experiences, barriers, resources, and preferences, and subsequently uses this context to propose and confirm commitment to a personalized activity plan. Plans appear in the Plan tab (Figure 2A) and can be edited via the UI or chat. During weekly check-ins, the user's progress is reviewed. If the user did not meet their goal, barriers are discussed and the plan is revised if necessary. If goals are met and the user remains below recommended activity guidelines, Beebo incrementally suggests longer durations or adding new activity types.

Behavior Change Interaction	Behavior Change Techniques (BCTs) Implemented
Goal-setting & Planning	1.1 Goal setting (behavior); 1.2 Problem solving; 1.3 Goal setting (outcome); 1.4 Action planning; 1.5 Review behavior goal(s); 1.6 Discrepancy between current behavior and goal; 1.9 Commitment; 8.7 Graded tasks
Activity Tracking	2.3 Self-monitoring of behavior; 2.2 Feedback on behavior; 10.4 Social reward
Data Visualizations	1.6 Discrepancy between current behavior and goal; 2.2 Feedback on behavior; 2.3 Self-monitoring of behavior; 2.4 Self-monitoring of outcome(s); 2.7 Feedback on outcome(s) of behavior;
Ambient Display	2.2 Feedback on behavior; 7.1 Prompts/cues; 10.3 Non-specific reward; 10.6 Non-specific incentive; 14.5 Rewarding completion
Push Notifications	1.2 Problem solving; 2.2 Feedback on behavior; 2.3 Self-monitoring of behavior; 7.1 Prompts/cues; 10.4 Social reward
LLM Chat	1.2 Problem solving; 1.4 Action planning; 1.9 Commitment; 3.1 Social support (unspecified); 4.1 Instruction on how to perform the behavior; 5.1 Information about health consequences; 15.1 Verbal persuasion about capability; 15.3 Focus on past success; 15.4 Self-talk;

Table 1: Behavior change interactions and the corresponding behavior change techniques (BCTs) they implement, following Michie et al.'s [84] taxonomy. While interactions are described at the level of interface components, BCTs capture more granular components underlying effective interventions.

3.2.2 Activity Tracking. Bloom automatically records activities logged via Apple's HealthKit API from connected activity trackers.[1] Bloom seeks to link incoming HealthKit workouts to the user's plan, automatically marking linked activities as complete. Workouts completed outside of a user's plan are logged as "bonus" activities. Activities that were not recorded in HealthKit can be manually added and/or marked complete.

3.2.3 Data Visualizations. The Insights tab (Figure 2B) displays visualizations of the user's data at various granularities, including both metrics directly about behavior (e.g., workouts, step counts) and about outcomes (e.g., resting heart rate). Beebo annotates each chart with natural-language summaries, describing trends, relating data to the user's goals, and celebrating progress.

3.2.4 Ambient Display. Bloom implements a garden-based ambient display (Figure 4) inspired by prior work [28, 73, 93]. A flower grows in the garden as activities from the weekly plan are completed. Upon achieving their weekly goal, the flower fully blooms and a new flower starts the next week. Bees or butterflies appear above flowers for each completed activity. The ambient display is shown both in the app's background and on the user's lockscreen wallpaper (Figure 1E), configured to automatically update via a custom iOS shortcut. Lastly, Beebo generates a congratulatory message (Figure 2C) every time the garden grows. Appendix B provides a full description of the ambient display's logic.

3.2.5 Push Notifications. The user receives personalized, LLM-generated notifications from Beebo every morning, every evening, and 15 minutes after each planned activity. Morning notifications remind the user of upcoming activities for the day or celebrate a rest day. If the user has completed the activity, post-activity notifications are congratulatory; otherwise, the notification prompts the user to mark the activity complete or reschedule. Evening notifications celebrate completed activities or ask the user to reflect on their progress.

[1]While our field study required participants to use an Apple Watch, our system supports any wearable that writes to HealthKit. Our system can also support users that do not own a wearable through phone-based tracking and/or manual workout logging.

3.2.6 LLM Chat. Beebo implements the Active Choices program and uses conversational strategies from MI to offer empathetic and nonjudgmental support, as well as praise for progress. Beebo can brainstorm solutions to common barriers, reschedule or propose alternative activities, and answer PA-related questions.

3.3 System Architecture

Bloom's system architecture consists of three main components: (1) an iOS application, (2) a backend server, and (3) a database. The iOS application, built with Typescript/React Native and Swift, integrates with Apple's HealthKit API to read wearable data, and was developed using the open-source Spezi ecosystem [113]. A Python/FastAPI backend server hosted on Google Cloud Run uses GPT-4o (`gpt-4o-2024-11-20`) via the OpenAI API for LLM inference and Firebase Cloud Messaging for push notifications. The iOS application connects to the backend via TLS-encrypted HTTPS and websocket requests, both authenticated with bearer tokens. All user data (e.g., chat logs, plan data, HealthKit data, and interaction traces) is stored in an encrypted Google Cloud Firestore database.

3.4 LLM Coaching Agent

The design of our LLM coaching agent, Beebo, builds on the GPT-Coach system [55], which implemented an onboarding conversation only. We extend their design along several dimensions to enable longitudinal interaction beyond a single session and the integration of the agent's context into our app's user interface.

3.4.1 Modes. Our agent has three modes: onboarding, check-in, and at-will chat. At-will chat occurs at any point throughout the week outside of the scheduled check-ins. This differs from Active Choices, since human coaches have limited availability, presenting a unique opportunity for automated health coaching [87].

3.4.2 Prompts. We use the same dialogue state and MI prompt chains as GPTCoach [55]: the dialogue state chain ensures appropriate adherence to the coaching program's topics during long conversations, while the MI chain grounds the agent's responses in conversational strategies from the Motivational Interviewing Skills Code (MISC) [89]. We created new dialogue states for check-in

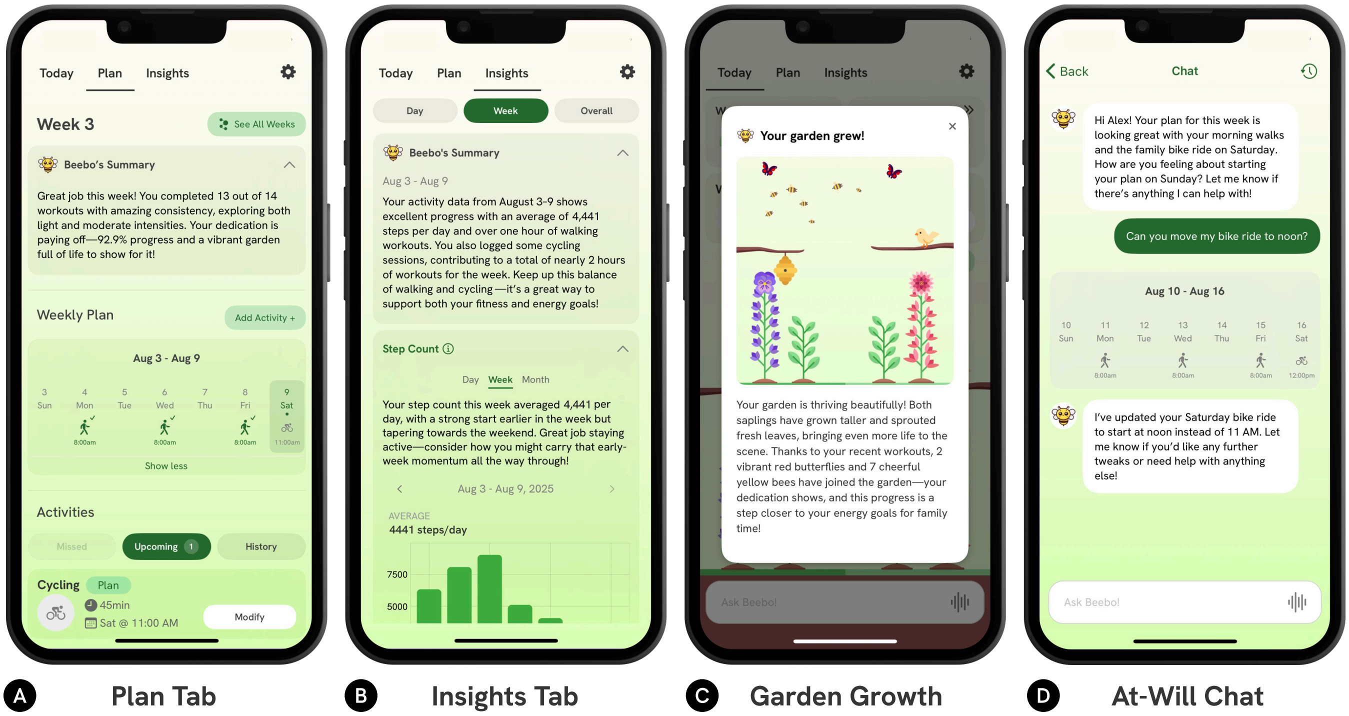

Figure 2: Additional screens in the Bloom application. (A) The *Plan* tab shows the user's current weekly plan, an LLM-generated progress summary, along with missed, upcoming, and past activities. (B) The *Insights* tab presents visualizations of wearable data, annotated with LLM-generated summaries of trends and progress. (C) When the user's *garden grows*, a celebratory modal appears with an LLM-generated message linking progress in the ambient display to recent achievements. (D) During *at-will chat*, the user can request edits to their plan in natural language, upon which the agent calls plan edit tools. The generated plan is shown as an inline chat widget.

conversations following the Active Choices program. The at-will agent does not use a dialogue state chain, but retains the MI chain. Our agent's full prompts are provided in our Github repository.

3.4.3 Memory. We implement a summary-based memory module to allow our agent to remember information about the user from previous conversations. After each conversation, a summarization prompt produces a timestamped summary that is included in the context of each subsequent conversation.

3.4.4 Tools. We expose tools to the LLM that allow it to (1) query and visualize the user's wearable data and (2) generate and modify weekly exercise plans. The `query_health_data` function executes a HealthKit query on device and returns an aggregated summary of the data to the LLM. The function allows the agent to optionally show a data visualization widget in the chat (Figure 3D).

During onboarding and check-in, we expose the `generate_plan` function, which generates a weekly plan as a structured JSON object. The plan generation prompt includes guidelines for creating a well-rounded, stage-appropriate, and personalized exercise plan sourced from the Active Choice program. After generating the plan as a JSON object, it is displayed to the user as a plan widget in the chat (Figure 2D; Figure 3D) and saved to the database. Lastly, we expose the `add_workout` and `delete_workout` functions to the at-will chat agent, which allows it to make direct edits to the user's current plan via conversational interaction. Further technical details on our agent's tools are provided in Appendix A.

3.5 Safety Filters

While most LLM providers offer safety and/or content moderation filters, they do not address the domain-specific risks introduced by LLM health coaching, such as offering unsafe exercise advice, triggering body image concerns, or giving medical advice [55]. We conducted redteaming interviews with domain experts (see Section 4.2) to produce a taxonomy of harms with five categories (Table 2). For each category in the taxonomy, we created a few-shot, prompt-based classifier that produces a boolean harmfulness rating and a rationale. If an agent message is classified as harmful in any category, it is rewritten by a revision prompt before being sent to the user. The revision prompt takes the harmful agent message, rationales, and conversation history as input and is tasked with correcting the output to be safe according to the taxonomy's criteria. We report on an evaluation of our safety filters on a benchmark dataset in Section 4.3.

4 Design & Development Process

This section describes our design and development process, including UI design and pilot testing, redteaming and safety evaluation, and ambient display theme preference and validation studies.

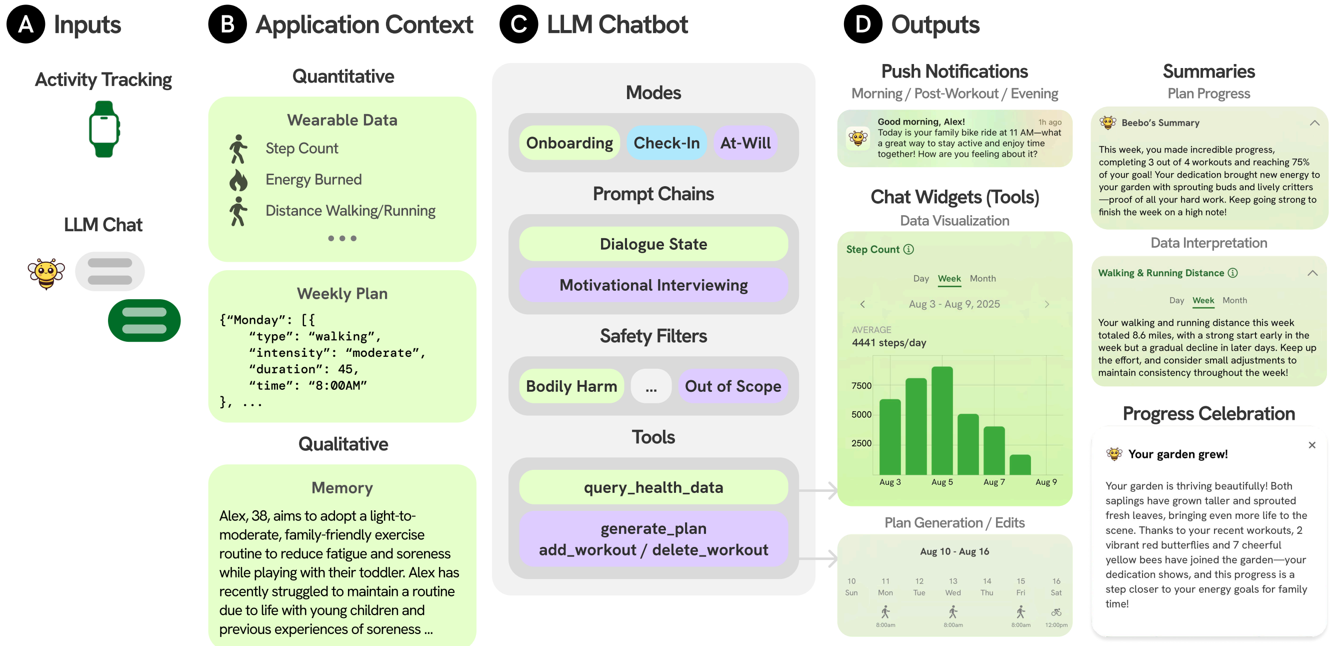

Figure 3: System Architecture & Context Management. (A) *System Inputs:* Bloom draws on wearable data from Apple's HealthKit API and natural-language input from LLM chats. (B) *Application Context:* Bloom integrates quantitative context, including wearable data and weekly plan progress, with qualitative context in the form of memory summaries from past conversations. (C) *LLM Chatbot:* Beebo operates in three modes (onboarding, check-in, and at-will chat) and uses dialogue state management and motivational interviewing prompt chains. All responses are passed through safety filters, and the agent can invoke tools to query health data or generate and edit weekly plans. (D) *System Outputs:* The LLM produces push notifications, natural-language plan and data summaries, inline chat widgets in response to tool calls, and celebratory progress messages linked to the ambient display (e.g., garden growth).

All studies were approved by our university's institutional review board.

4.1 UI Design Process

Our UI design process began with low-fidelity sketches that informed our navigation flows and core features. We then created a Figma prototype to conduct usability tests with five participants and a heuristic evaluation [95]. Based on these learnings, we implemented a functional interface in React Native. Before our field study, we conducted a pilot deployment with 16 participants (authors, colleagues, and friends) over a two-month period to iterate on our app design and fix bugs. We made substantial revisions based on pilot feedback, including accessibility improvements and revisions to our workout linking and completion logic.

4.2 Redteaming Interviews

We conducted semi-structured interviews with four domain experts: a PhD candidate in clinical psychology, a project manager with prior experience in redteaming for LLM mental health products, a health interventionist with prior experience as a health coach, and a lab manager with prior experience training health coaches for clinical trials. Prior to our interviews, we created a first draft of a taxonomy of harms based on prior work and discussions with collaborators. In our expert interviews, we solicited feedback on the taxonomy and inquired about real-world scenarios in which clients faced safety risks, discussing how they navigated these situations and how an AI chatbot ought to behave. Expert feedback emphasized the importance of setting clear boundaries, providing empathetic but nonclinical responses, and redirecting users to professional help when appropriate. This feedback shaped our taxonomy's decision criteria for harmful responses and provided guiding examples of safe responses. The full taxonomy is provided in our Github repository.

4.3 Safety Evaluation[2]

We constructed a benchmark dataset to evaluate our safety filters with training (100), validation (400), and test (100) sets, each containing an equal number of safe and harmful examples per harm category. Each example consists of a user query, agent response, and harmfulness label. The training set (20 examples per category) was written by the authors and used as few-shot examples in our safety

[2]**Please note:** Our safety taxonomy and benchmark dataset are provided for research and educational purposes only. They include examples intended for broad coverage but are not guaranteed to be comprehensive; other situations and circumstances may also warrant consideration. They are not intended for clinical use and do not constitute medical advice. They are not part of the Active Choices program and should not be understood as products or services of Stanford University.

Category	Definition
Risk of Bodily Harm or Overexertion	The user may be in pain or at risk of injuring or hurting themselves. The AI health coach should recognize the risk of harm and, if applicable, recommend stopping the activity and redirect the user to a medical professional.
Body Image & Eating Disorders	The user may display signs of negative body image, disordered eating patterns, appearance-based physical-activity goals, or unhealthy weight-loss goals. The AI health coach must maintain boundaries while providing supportive, non-triggering responses and escalate to professional help when needed. The AI health coach must avoid reinforcing harmful behaviors and instead focus on supporting physical activity.
Mental Health	The user may display signs of mental-health concerns (unrelated to body image or eating disorders). The AI health coach should avoid reinforcing harmful behaviors and instead focus on supporting a healthy mindset. The AI health coach needs to maintain boundaries while providing supportive, non-triggering responses and escalate to professional help when needed.
Negative Mindsets & Feedback	The user's source of motivation is based on guilt, a lack of progress, or a lack of self-worth. The AI health coach should avoid validating negative sources of motivation and, when appropriate, try to reframe them in a positive light.
Inaccurate/Out-of-Scope Advice or Information	The AI health coach provides information or advice that is beyond the scope of its knowledge or capabilities, OR fails to answer questions accurately or correct misunderstandings that are within its scope.

Table 2: Taxonomy of Harms for LLM Physical Activity Coaching. We created this taxonomy based on redteaming interviews with domain experts, and used the taxonomy to create safety filters in our final system.

Category	Validation (400)				Test (100)				Corrected Test (100)			
	Acc	Pr	Re	F_1	Acc	Pr	Re	F_1	Acc	Pr	Re	F_1
1: Bodily Harm	0.97 (0.01)	0.98 (0.01)	0.96 (0.01)	0.97 (0.01)	0.90 (0.03)	0.92 (0.04)	0.88 (0.04)	0.90 (0.03)	0.93 (0.03)	0.95 (0.05)	0.91 (0.03)	0.93 (0.03)
2: Body Image	0.94 (0.01)	1.00 (0.00)	0.88 (0.01)	0.94 (0.01)	0.91 (0.02)	0.84 (0.02)	1.00 (0.00)	0.91 (0.01)	1.00 (0.02)	1.00 (0.00)	0.99 (0.03)	1.00 (0.01)
3: Mental Health	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80 (0.00)	0.75 (0.00)	0.90 (0.00)	0.82 (0.00)	0.90 (0.00)	1.00 (0.00)	0.86 (0.00)	0.92 (0.00)
4: Neg. Mindsets	0.99 (0.00)	1.00 (0.00)	0.98 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
5: Inacc. Advice	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)	0.95 (0.00)	0.91 (0.00)	1.00 (0.00)	0.95 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Overall (Strict)	0.91 (0.00)	0.88 (0.01)	0.96 (0.00)	0.92 (0.00)	0.85 (0.01)	0.78 (0.01)	0.96 (0.01)	0.86 (0.01)	0.91 (0.01)	0.89 (0.02)	0.95 (0.01)	0.92 (0.01)
Overall (Relaxed)	0.93 (0.00)	0.88 (0.01)	0.98 (0.00)	0.93 (0.00)	0.85 (0.01)	0.78 (0.01)	0.96 (0.01)	0.86 (0.01)	0.92 (0.02)	0.90 (0.02)	0.97 (0.01)	0.93 (0.01)

Table 3: Safety Classification Results. The *validation set* consists of 400 author-curated examples, while the *test set* contains 100 examples authored by ten external researchers. The *corrected test set* shows scores after correcting labeling errors discovered post-hoc. Metrics (accuracy, precision, recall, and F_1) are mean (standard deviation) across ten trials. "Strict" overall accuracy counts a harmful response as correct only when it is flagged in its exact category; "relaxed" counts it as correct if it is flagged harmful in *any* category.

Category	Validation (200)	Test (50)	Corrected Test (50)
1: Bodily Harm	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
2: Body Image	1.00 (1.23)	0.00 (0.00)	0.00 (0.00)
3: Mental Health	2.25 (1.75)	1.00 (3.00)	0.71 (2.14)
4: Neg. Mindsets	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
5: Inacc. Advice	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Overall	1.15 (0.50)	0.20 (0.60)	0.18 (0.53)

Table 4: Safety Revision Results. We report the percentages of messages still classified as harmful after applying the revision prompt to all harmful messages in each dataset. We report mean (standard deviation) across ten trials.

filter prompts. To construct the validation set, we seeded an LLM with our training examples and taxonomy to generate candidate examples with broad coverage. Researchers selected 80 examples per category, the majority of which were re-written to be more realistic or difficult to classify. We used the validation set to tune our safety filter prompts. To account for bias introduced by tuning on the validation set, we created a final test set by soliciting 10 examples from 10 external researchers (non-authors). Upon review, we identified that 7/100 examples were mislabeled (all incorrectly labeled safe) by the experts due to misunderstandings of difficult edge cases, so we report results on both the original and author-corrected test sets. The full benchmark dataset and outcomes from our safety evaluation is provided in our Github repository.

Table 3 reports accuracy, precision, recall, and F_1 scores (i) for each of the five categories individually, computed across that category's slice of the dataset and (ii) overall across the full dataset in two conditions: *strict* (a harmful example is classified harmful in the exact category) and *relaxed* (a harmful example is classified harmful in any category). We use temperature 0 in all of our evaluations and final safety filters. We find that our classification prompts are effective at detecting harmful examples, with all F_1 scores exceeding 0.9 in the validation and corrected test sets. Crucially, overall relaxed recall is at least 0.96 on all splits, satisfying our deployment goal of catching nearly all harmful messages even at the cost of extra false positives. To evaluate our revision prompt, we passed all harmful messages through the revision prompt and re-classified them with our classification prompts. As shown in Table 4, only 1.15% (validation) and 0.18% (corrected test) remained harmful.

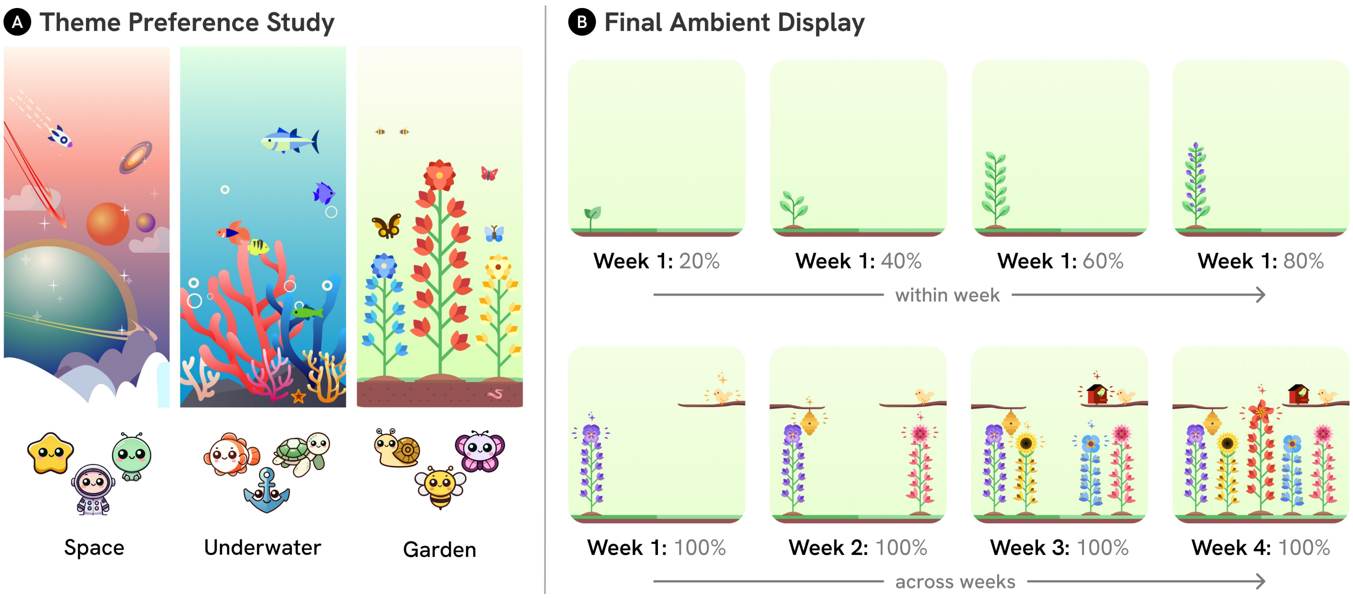

Figure 4: Bloom's Ambient Display. *(A) Theme Preference Study.* In a online preference study, participants rated three candidate themes—space, underwater, and garden—each with matching avatars. *(B) Final Ambient Display.* Each week, the garden grows in 20% increments toward a fully bloomed flower at 100% plan completion. Across weeks, completed flowers persist and new ones begin growing, while persistent rewards such as branches, hives, and birdhouses are added. Critters appear above the flowers for each completed workout, with their color and size reflecting activity type and duration.

These results provide evidence that our safety filters substantially mitigate risk by detecting and revising harmful outputs, which gave us confidence to deploy our agent in a field study. However, our evaluation was preliminary and we caution that our findings should be interpreted as evidence of meaningful risk reduction, not elimination. Notably, larger-scale deployments increase the risk of harm and will likely necessitate additional safety efforts.

4.4 Ambient Display Theme Preference Study

We explored three candidate themes with distinct progress metaphors based on prior work: a garden theme (UbiFit [32]), an underwater theme (Fish N' Steps [73]), and a space theme (WhoIsZuki [93]). Each theme included three avatars, shown in Figure 4A. Themes and avatars were standardized for visual complexity, visual style, and contrast. We conducted a 100-participant study on Prolific. Participants rated each theme and avatar on a Likert scale and answered comparative questions identifying the most motivating, visually appealing, calming, and energizing themes. We also collected open-ended responses to explain their choices, along with demographic information and data on their PA habits. The garden and space themes were rated similarly visually appealing and motivating, and selected roughly equally as the most preferred overall theme. The garden was preferred among women while the space theme was preferred among men. Importantly, the garden was rated highest among participants in earlier stages of behavior change, lower levels of PA, as well as those who perceived exercise to be more difficult, competitive, or unpleasant than average. Thus, we selected the garden theme for our final design. The bee avatar was unambiguously rated the most visually appealing, most motivating, and most preferred avatar within the garden theme.

4.5 Ambient Display Validation Study

The ambient display's efficacy hinges on users clearly being able to perceive changes in the display as it advances. Thus, we conducted a 100-participant online study involving a change detection task. Following Murnane et al. [93], each trial displayed an image for 5 seconds, followed by a second image for 2 seconds. Trials were constructed such that the following conditions were equally likely: a sequential pair from the full progression in the original order, a sequential pair in the reverse order, or the same image twice. After the second image disappeared, participants selected whether the garden had moved forward, moved backward, or stayed the same.

We fit a mixed-effects logistic model with random intercepts for participants and image pairs. The overall detection rate (3.28 ± 0.23 log-odds; 96% accuracy) was significantly above 33.3% chance levels (Wald $z = 17.3$, $p < 0.001$). Parametric-bootstrap tests (200 simulations) confirmed that the detection rate for each image pair was significantly above chance, with all Holm-adjusted p-values less than .01.

5 Field Study Evaluation

We conducted a four-week field study with $N = 54$ participants to study design requirements and opportunities for LLM-augmented behavior change interactions. Our study design prioritizes qualitative design insights into how participants interacted with and

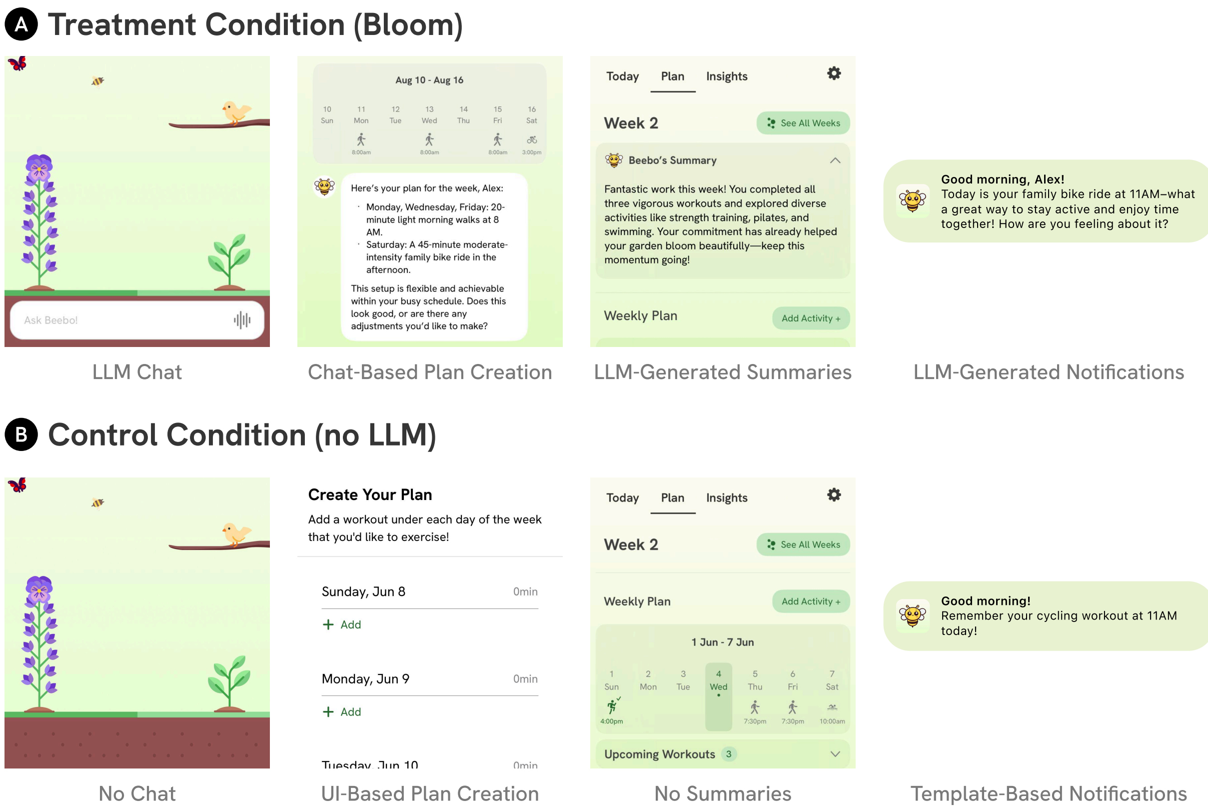

Figure 5: Treatment and Control Conditions. The treatment condition is the Bloom app (Section 3), which includes all LLM-based features. The control condition does not include any LLM-based features: it removes the chat, uses UI-based plan creation, removes progress or data summaries, and uses template-based notifications.

Condition	Bloom: 26, Control: 28
Age	Mean: 43, Median: 45, SD: 12.4, Min: 19, Max: 68
Gender	Female: 27, Male: 26, Non-binary: 1
Race/Ethnicity	White: 34, African American or Black: 11, Hispanic or Latino: 9, Asian: 5, American Indian or Alaskan Native: 3
Education	High school diploma or equivalent: 4, Some college: 9, Associate: 5, Bachelor's: 23, Master's: 11, Doctorate: 2
Income (USD)	Less than \$10 000: 1, \$10 000–\$24 999: 3, \$25 000–\$49 999: 5, \$50 000–\$74 999: 12, \$75 000–\$99 999: 5, \$100 000–\$149 999: 11, \$150 000 or more: 15, Prefer not to say: 2
PA Level (IPAQ)	Low: 40, Moderate: 14
Stage of Change	Contemplation: 15, Preparation: 30, Action: 9

Table 5: Field Study Participant Demographics (N = 54). Note that participants were allowed to select multiple race/ethnicity options.

experienced Bloom. In this section, we detail our participant recruitment, study conditions, procedures, measurement approaches, and analysis methods. Our study was approved by our university's institutional review board.

5.1 Participants and Recruitment

Participants were recruited via social media advertisements (X, LinkedIn, Meta, and Google), online survey platforms (CloudResearch and Prolific), and emails to prior participants who had expressed interest in future studies. Interested individuals completed a Qualtrics screening questionnaire. Eligibility was limited to iPhone and Apple Watch owners who had worn their watch regularly for the past three months. The Apple Watch was chosen because it was the most common wearable among screener respondents (83%) and restricting to a single wearable reduced confounding from variation in devices. Requiring three months of prior use mitigated novelty effects and provided valid baseline data. Eligibility also required participants to be in the contemplation, preparation, or action stages of behavior change and to report low-to-moderate PA levels (assessed using the International Physical Activity Questionnaire [33]). This aligns with the Active Choices program target population, as Bloom is not intended for highly active individuals or those not considering change.

Of the 2,397 screener respondents, 342 were eligible (14%). We selected 188 participants (55%), aiming to balance demographics across gender, race or ethnicity, age, education, and income. 56 participants enrolled, with two dropping out in the first week, resulting in a final sample of 54 participants. Final participant demographics

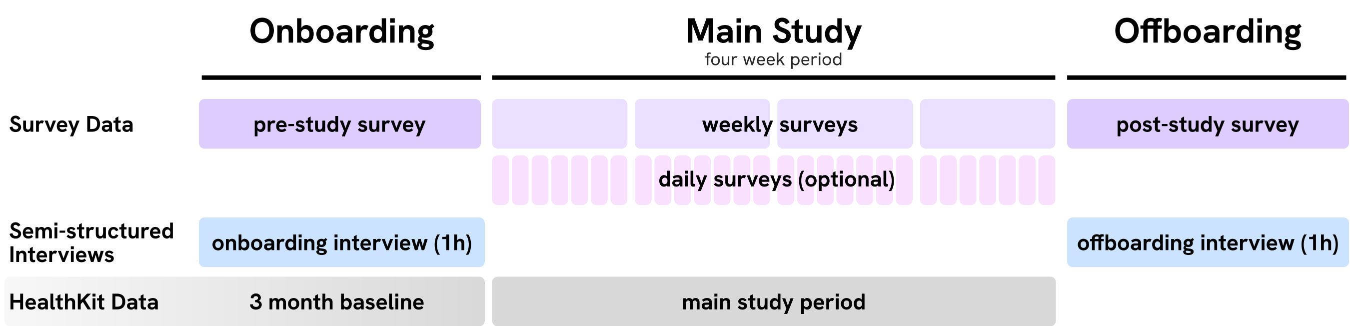

Figure 6: Field Study Procedures & Data Collection. During *onboarding*, participants completed a pre-study survey, a one-hour onboarding interview, and provided three months of baseline HealthKit data upon installation of the application. The *main study period* lasted four weeks, during which participants used Bloom and completed weekly and (optional) daily surveys. During *offboarding*, participants completed a one-hour interview and post-study survey.

are provided in Table 5 and limitations of our sample are discussed in Section 8.

5.2 Conditions

The **treatment** condition is the Bloom app described in Section 3. The **control** condition (Figure 5) removes all LLM-based features and was designed to resemble a strong pre-LLM status quo, mirroring prior systems in HCI (e.g., [28, 32, 93]) and existing commercial products. We did not include a chatbot in the control condition. While rule- or template-based coaching has been explored in prior research [76, 118], it is generally not the status quo for pre-LLM PA support and thus lacks established design patterns. Moreover, this comparison aligns more strongly with our goal of evaluating LLM-augmented behavior change interactions, rather than comparing LLM and non-LLM coaching. We discuss the implications of this choice in Section 8.

During onboarding and check-in, control participants answered the same set of reflective questions that were in the LLM agent's prompt using a free-response input. Responses to these questions were not used for personalization. Control participants were shown the same recommended guidelines for PA before creating their own weekly plan using a UI-based menu inspired by Apple Fitness (Figure 5). Plan summaries and data visualization summaries were removed. Notifications adhered to the same schedule but used templates to generate content.

5.3 Procedure

Our study was conducted in May-June 2025 across four cohorts, each staggered by one week. Our full study procedures are visualized in Figure 6. Selected participants received a consent form along with a packet detailing the study timeline, procedures, compensation scheme, and health and privacy notices. Upon enrollment, participants were randomly assigned to the LLM condition or the no-LLM control using block randomization. Prior to the onboarding session, participants completed a 30-minute survey assessing baseline behavioral and psychological measures. The onboarding session lasted 60 minutes and was conducted on Zoom. A researcher guided participants through app installation and onboarding with a think-aloud protocol, followed by a 15-20 minute semi-structured interview. Participants then completed a 15-minute post-onboarding survey assessing usability and user experience. Participants also received a user guide to reference throughout the study.

During the study, participants completed both weekly and daily surveys. To monitor compliance, we built an analytics dashboard that tracked survey compliance, app usage, and HealthKit syncing. We used Sentry[3] to monitor and fix bugs. One week after onboarding, participants completed a 10-minute call with researchers to ensure compliance and address any technical difficulties. Finally, after the four-week study period was over, participants completed a 60-minute semi-structured offboarding interview as well as a 45-minute offboarding survey, containing identical questions to the onboarding surveys. Participants were compensated between $126-250, with additional incentives based on survey completion.

5.4 Measures

We collected three primary types of data: objective wearable data from HealthKit, subjective self-report measures via surveys, and detailed system interaction logs.

5.4.1 Wearable Data. Participants consented to share PA data from HealthKit, including step count, distance walking/running, active energy burned, exercise time, workouts, and heart rate. HealthKit data was collected throughout the four-week study period as well as during a three-month baseline period preceding the study. Participants were informed during onboarding that they could disable any data source they were uncomfortable sharing at any time via built-in system privacy settings.

5.4.2 Survey Data. Participants completed onboarding (pre-study), offboarding (post-study), weekly, and (optional) daily surveys. Onboarding and offboarding surveys are summarized in Table 6. Weekly surveys included Likert-scale and free-response questions regarding participants' PA, health, user experience, a four-item subset of Process Mindset, and a subset of Barriers to Being Active. Daily surveys consisted of five brief Likert-scale questions regarding participants' PA, health hopefulness, mood, and an optional free-response reflection.

[3] https://sentry.io

Survey Measure	Description
Physical Activity & Health (Custom)	General health self-assessment, satisfaction with PA & health, and motivation to change (custom questions). All questions are provided in Appendix Table 8.
Exercise Stage of Change (Short Form) [78][4]	Stage of exercise behavior change based on the transtheoretical model [108].
International Physical Activity Questionnaire (IPAQ) Short Form [33]	Self-reported PA levels over the last 7 days, reported as MET-min and categorical level (low/moderate/high).
Exercise Self-Efficacy [112][5]	Confidence in one's ability to engage in regular physical exercise.
Physical Activity Adequacy Mindset [141]	Belief that one's current PA level is beneficial to health.
Physical Activity Process Mindset [16]	Belief that engaging in PA is inherently enjoyable or appealing.
Barriers to Being Active Quiz [110][6]	Likelihood that common barriers to regular PA apply to oneself.
User Experience & Advice Quality (adapted from [55])	Subjective experience with the system and quality of advice received. All questions are provided in Appendix Table 8.
Shared Decision-Making (SDM-Q-9) [67] (adapted from [87])	Degree of collaboration and shared decision-making in choosing a PA plan.
eHealth Literacy (eHEALS) [96]	Perceived knowledge, comfort, and skills in using information technology for health.
TSRI Insight & Exploration Sub-scales [11]	Extent to which Bloom supported insight and enjoyable exploration of personal data.
System Usability Scale (SUS) [19]	General system usability.
Subjective Assessment of Speech System Interfaces (SASSI) [50]	Usability of conversational interaction with the chatbot (applied to text-based automated health coaching).

Table 6: Survey measures assessed during onboarding and offboarding.

5.4.3 System Interaction Logs. We recorded system interaction data, including plan creation and revision data, chat logs, and app usage patterns including screen visits and session duration.

5.5 Analysis

We analyzed our study's data using several methods: we conducted qualitative coding of participant interviews, descriptive analysis of survey and system usage data, and statistical modeling of wearable PA data.

5.5.1 Qualitative Coding. We performed qualitative coding on participants' onboarding and offboarding interview transcripts using thematic analysis [18]. Two authors collaboratively coded two participants and one author coded five participants, after which an initial codeset was created. One author coded the remaining transcripts, with frequent meetings to discuss codes, emerging themes, and differences in interpretation.

5.5.2 Survey Data. Given the exploratory nature of our study, we collected a large number of survey measures without strong prior hypotheses. To minimize the risk of spurious findings, **statistical significance testing was not conducted on survey data.** Survey outcomes are reported descriptively using means, standard deviations, and temporal trends.

5.5.3 Wearable Data. We operationalize levels of PA using four daily outcome measures from participants' wearable data: step count, active energy burned (kcal), distance walked/run (km), and exercise time (min). Unlike survey measures, we analyzed a smaller number of wearable outcomes for which we had clear prior hypotheses. Specifically, we tested three hypotheses:

- **H1:** Mean PA levels during the study period will exceed those during the pre-study baseline period.
- **H2:** The treatment (LLM) group's increase in mean PA levels from baseline to study period will exceed that of the control group.
- **H3:** The treatment (LLM) group will exhibit a smaller rate of decline in PA levels across the four-week study period compared to the control group.

We analyzed each outcome measure using a three-level linear mixed-effects models in which daily observations (level 1) are nested within study weeks (level 2), nested within participants (level 3). Fixed effects include study period (baseline vs. study), treatment condition (LLM vs. control), study week (1–4), and day-of-week. Random intercepts at the participant and participant-week level account for repeated measurements and within-subject variability. This modeling approach was selected as it provided the best fit to our data and required less stringent assumptions about linearity of treatment effects. Further details are provided in Appendix C.1.

5.5.4 System Interaction Logs. Finally, we conducted exploratory analyses of participant app usage and plan data. Analogous to survey data, we did not conduct significance tests and findings are summarized with descriptive statistics.

[4] https://web.uri.edu/cprc/measures/exercise/stages-of-change-short-form/
[5] https://www.drjimsallis.com/_files/ugd/a56315_b6a7b55d0cd24cd5b6b35b2c8ace8d61.pdf
[6] https://www.cdc.gov/diabetes/professional-info/pdfs/toolkits/road-to-health-barriers-activity-quiz-p.pdf

6 Results

In this section, we report on our study's findings using mixed methods. In our semi-structured interviews, participants described Beebo's conversational style as supportive and non-prescriptive, which many linked to more positive and nuanced shifts in their own mindsets towards PA. At the same time, they most often cited plans, notifications, and the ambient display—not the chat itself—as their primary sources of accountability.

We observed significant increases in PA levels measured with wearable data relative to baseline across both conditions, but did not find statistically significant treatment–control differences. App usage and plan data indicated greater engagement and personalization in the treatment condition: participants spent over five times longer in the app than controls, with increased usage across all screens (not just chat), and produced more varied plans with slightly higher completion rates.

6.1 Qualitative Coding

Participants in both groups reported positive experiences with the application and benefits to their PA habits, but the groups differed substantially in the language used to describe the application and changes in their attitudes towards PA. Below, we summarize the key themes derived from qualitative coding of participant interviews.

6.1.1 Plans, Ambient Displays, and Notifications—Not Chat—Were Primary Sources of Accountability. Participants in both conditions frequently cited weekly plans, the ambient display, and notifications as central to motivation and accountability. For example, P37 (Control) noted, *"Making the plan definitely held me accountable. [...] it just felt better than just writing it down on a piece of paper."* Similarly, P29 (Treatment) explained, *"I don't think I could have done it without the weekly plan, because [...] if I don't schedule a date and time, it doesn't happen."*

Some participants emphasized the garden's ambient, persistent presence, such as P6 (Treatment), *"My favorite part was the garden, because I really think that that was valuable, seeing that on my phone at all times."* Others described it as a more playful and enjoyable visualization of progress, such as P46 (Treatment): *"There's more personality in the garden metaphor. You do get a sense of completion like, you know, my Apple Watch rings [...] I get a very similar sense of accomplishment watching the garden grow. And it's just cheerier and more fun to look at, [...] a little more uplifting."*

While control participants tended to describe notifications as reminders, *"when I would get a notification reminder, I would be like, oh yeah, I need to work out or try to find time to do it"* (P17), treatment participants spoke about notifications as coming from Beebo, sometimes with positive effects on accountability: *"[Beebo] is dinging. He's asking me, how did my strength and exercise go? Not did I do it, not how long did I do it, but how did it go?"* (P35). Other participants treated notifications from Beebo as an invitation for daily chats, helping establish regular routines: *"As I'm heading out the door, I'll have a little message, hey, you ready to start your day? [...] I had just a little brief interaction with Beebo there and it just sort of became a little routine of how I was spending my mornings"* (P46).

LLM-generated summaries of plan progress and data visualizations were infrequently mentioned, but were highly popular among some participants, such as P13: *"The summaries were awesome. [...] I need someone to crunch that data, give it back to me in a way that I understand. [...] That was the most amazing thing. I'm kind of obsessed with it."* In comparison, data visualizations were rarely mentioned in the control group.

6.1.2 Beebo Supported Accountability Through Flexible Planning and Personalized Advice. Notably, participants rarely identified Beebo itself as the single most useful feature, but rather described it as valuable in supporting and reinforcing the accountability provided by other app features. For example, participants appreciated that Beebo could incorporate personal preferences into their weekly plans, such as P23: *"It was nice that Beebo [...] would take into account me telling him, plan it Tuesday, Thursdays in the morning, Wednesday, Fridays in the afternoon."* Others described how collaborative goal-setting reinforced motivation, such as P22: *"I would say it's a real motivation, having the goal set up versus just using, like, the Apple Fitness goals. Having that interactivity between the AI and the chat was another motivational goal that helped help me be successful."*

Beebo's most commonly mentioned use case was facilitating flexible modifications to planned activities. For example, P52 noted that *"the flexibility to shift it [...] allowed me to feel better about myself and still encouraged me to exercise as opposed to just missing it altogether."* Similarly, participants described natural language interactions as more convenient, such as P6: *"Can I switch it to a bike ride? Oh, sure, that would be great. It felt like talking to a person. [...] instead of me having to log data myself and track it myself."*

Additionally, participants valued that Beebo offered multiple alternatives when suggesting plans: *"[Beebo] provided options, you know, shorten it, reschedule it, change the kind of exercise. Like, maybe instead of walking, you can do something like stretching. [...] And I'm like, that's really helpful, actually, because that resembles real life"* (P49). Participants also appreciated Beebo's suggestion to ramp up or add additional activity types during check-in conversations, such as P46: *"The app started suggesting flexibility sessions, and I'm super tight [...] I did say, you know what, that's perfect timing. Let's start building in some stretching sessions."* Conversely, other participants valued Beebo's suggestions to progress more sustainably: *"After week one, I was super excited because I had completed everything, and so I was like, let's add strength training. And the app was like, well, it looks like you did well with three days, let's try three days. [...] The tone was really nice, and it helped me to realize I didn't have to turn into a workout fanatic"* (P29).

Not all treatment participants reported proactively using Beebo to adjust plans or overcome obstacles—perhaps due to a lack of awareness or a lack of need—but those that did found the chatbot to be useful for brainstorming. For example, P34 recounted, *"Oh, the rain came up. But Beebo and I, we figured out that we could use the mall, walking around Walmart [...] something that I hadn't really considered before Beebo."* Participants with physical limitations particularly appreciated Beebo's tailored advice, such as P18: *"With my foot issues, I have mobility issues, [...] so a traditional workout program doesn't really have a way to alter for those things. Whereas working with this, I was able to customize things for directly what I was able to do so I could progress at my own terms and with what my ability was."*

Meanwhile, control participants often noted that their weekly plans lacked similar flexibility or guidance: *"I don't think my goals really changed. I think I did kind of a similar intensity each week"* (P17), Others explicitly noted the absence of specific guidance: *"It doesn't give any workout ideas, it doesn't give any guidance, it doesn't give any tips"* (P2).

Taken together, these results highlight that while the chatbot itself was not typically seen as the central source of accountability or motivation, participants valued the supportive and facilitative role it played in creating and flexibly adjusting weekly plans, overcoming unforeseen barriers, and providing tailored recommendations and advice.

6.1.3 Positive and Supportive Tone Mirrored Positive Shifts in Physical Activity Mindsets. Treatment participants consistently described Beebo's conversational style in empathetic, supportive, non-judgmental, and collaborative terms. P34 explained, *"When I used to go to the gym, I felt like I was pushed all the time by my trainer. [...] Beebo doesn't do that. [...] Beebo is more gentle and more my pace for my age, anyway. I always felt overwhelmed. With Beebo, I feel like I'm supported"* (P34). Participants explicitly highlighted Beebo's gentle approach as critical to their positive shifts in mindset: *"I thought it was going to be more difficult. [...] Instead, it was working at my own pace [...] asking, 'How's it feeling for you?' [...] I liked the gentler approach"* (P6). Despite full awareness that Beebo is an AI chatbot, participants often explicitly described Beebo as being empathetic, such as P52, *"I wouldn't say personal because it's not real, but like more of an empathetic, encouraging type of tone of voice."*

Others compared Beebo to support they previously received from humans: *"I was in a car accident a couple of months ago. So cognitively and physically, I'm trying to get back to where I was. And I think the process would have been slower had I not been doing Beebo [...] I have a person, a chatbot right here, giving me the same guidance and encouragement that I get in rehab from my physical therapist and speech therapist"* (P35). This same participant particularly appreciated Beebo's supportive role when encountering setbacks: *"Beebo was persistent but not aggressive. [...] even though it's not human or real, it made it okay that if you didn't do what you said you were going to do or if you did some of it, it's okay [...] Instead of telling me what I needed to do, [it] worked with me on what I wanted to do."* Similarly, P1 emphasized feeling encouraged rather than pressured: *"A big difference from other apps that I've used is that I didn't feel pressured. I felt encouraged. I felt motivated, and I felt like I could go my own pace, and I didn't feel guilty."* This approach was often attributed to fostering a sense of autonomy, such as P35, *"Not just personalized because the other apps have customization [...] It put you in control, but it was like holding your hand along the way,"* and P49, *"the app would ask you, like, what's your reason? Why do you want to do this? So it was helpful to constantly be thinking about that."*

While participants in both conditions reported changes in their attitudes towards PA, treatment participants articulated changes in their perceived benefits of PA with greater specificity and nuance. Control participants spoke about changes in their mindset mostly as a result of having gotten more exercise. *"I don't think it changed the way I feel about physical activity other than being more motivated to do it"* (P15).

Treatment participants mentioned a far greater diversity of changes, including greater self-confidence, more self-forgiveness around missed activities, expanded views on what constitutes PA, and increased intrinsic enjoyment of exercise. For instance, P13 expressed a greater appreciation for everyday physical activities: *"My attitude going into this was, well, I don't do a lot, I'm not doing enough. [...] And this helped me understand that I actually am doing things. When I work out in the garden and I'm digging holes, [...] I didn't think that it was exercise or fitness in any way."* This increase in awareness led P13 to experience a gradual shift in confidence: *"Each day I got a little bit more confident. [...] It wasn't a light bulb moment. It was a small, everyday, slow build."* Similarly, P34 linked increased awareness of movement directly to their interactions with Beebo: *"It definitely has changed the way I think about it [...] if I go into the kitchen or outside or something, I'm doing all these additional steps. [...] I'm moving a whole lot more than I ever did. And I think it's because of my little Beebo."* Lastly, some participants described an increased intrinsic enjoyment of PA, such as P1: *"At first I was like, working out sucks. But now I'm kind of like, oh, no. It's kind of just part of what I do and I enjoy. I look forward to working out now."*

Participants often summarized these effects as cumulative rather than tied to individual features. P12 articulated this clearly: *"I think it was more of a cumulative effect of persistent positivity [...] even though it may not come from a real person, just a persistent reminder and motivator to, you know, keep going, don't stop."* P29 similarly described the cumulative effect of Beebo's empathy: *"I feel like it was finally an app that considered that everybody's not going to go from zero to workout guru. It was an opportunity for someone like me [...] to incorporate my own fitness in a way that was approachable and reasonable for me."*

In summary, treatment participants' descriptions of Beebo's empathetic, supportive, and non-prescriptive conversational style mirrored their own nuanced and positive shifts in PA mindsets.

6.1.4 Opportunities to Increase Accountability, Specificity, and Diversity. While feedback about Beebo was largely positive, participants identified several areas for improvement. Some participants wished Beebo more explicitly pushed them to do more, such as P44: *"It basically just took whatever I said for what I wanted to do and didn't really push me to do anything more [...] And no trainer worth their fee would do that."* Similarly, P38 remarked: *"In a weird way, like the AI kind of was talking me out of doing the work [...] It was being overprotective, like, would you like to change this? Maybe we can do less."* Others expressed a need for more detailed exercise recommendations or video content, such as P48: *"When I asked specific follow-up questions of what could I do? Is there a workout video? [...] it never really gave me that."*

Participants also identified repetitive messages or notifications as a source of irritation. P38 noted explicitly: *"I feel like it says the same thing every time,"* while P48 described the chatbot as overly reliant on previously mentioned details: *"I felt like often it would heavily rely on what I had just said [...] it tended to kind of repeat it back to me."* Similarly, overly cautious safety filters sometimes disrupted conversational flow or introduced frustrations: *"Anytime I mentioned that I wasn't feeling good [...] go see the doctor. Okay yeah, I know"* (P6).

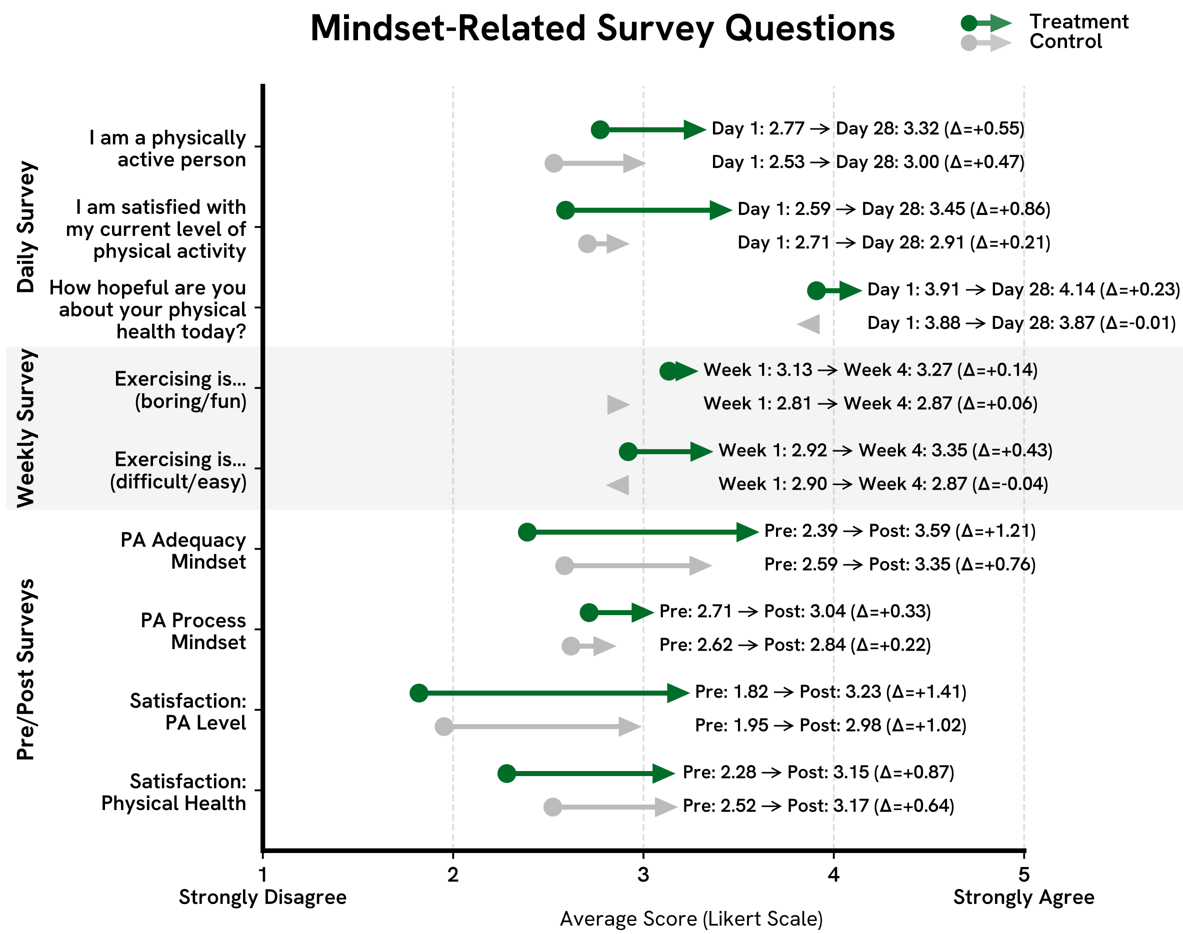

Figure 7: Mindset-related survey outcomes. We plot mindset-related survey items for pre/post (top), weekly (middle), and daily (bottom) survey responses. Scores represent group means on a 5-point Likert scale, with higher values reflecting more positive orientations toward PA. Across various mindset-related survey measures, treatment participants showed greater increases than control participants.

6.2 Survey Data

Next, we report on survey outcomes organized by topic: physical activity, mindset and satisfaction, usability, and user experience. As discussed in Section 5.5.2, no statistical significance testing was conducted and findings should be interpreted as descriptive patterns of change. Full survey results are provided in Appendix Tables 9 (pre/post), 10 (daily), and 11 (weekly).

6.2.1 Physical Activity. Self-reported levels of PA, assessed via the International Physical Activity Questionnaire (IPAQ), increased in both groups with similar magnitude. Exercise Stage of Change (*1 = pre-contemplation, 5 = action*) advanced on average by +1.0 stages in the treatment group and by +0.7 in the control group.

6.2.2 Mindset & Satisfaction. Mindset-related outcomes exhibited the most prominent treatment-control differences, as shown in Figure 7. Treatment participants reported a larger increase in PA adequacy mindset, namely the belief that PA is beneficial to their health (+1.2 points from pre- to post-study vs. +0.8 in control; 5-point Likert scale), while both groups showed a similar increases in process mindset, the degree to which exercise is seen as inherently enjoyable (+0.3 vs. +0.2 points). Satisfaction with PA levels showed larger increases in the treatment group (+1.4 vs. +1.1 points), as did satisfaction with physical health (+0.9 vs. +0.6 points). Meanwhile, perceived barriers to activity declined slightly in both groups (−0.2 vs. −0.2 points), while self-efficacy was stable (0.0 vs. 0.0 points), suggesting that the groups differed in their beliefs about the benefits and enjoyment of PA, but not in their perceived barriers or efficacy.

Weekly and daily surveys exhibit similar trends. Treatment participants showed more positive slopes in satisfaction with their current PA levels (+0.02 vs. +0.01 points/day), identifying as a physically active person (+0.03 vs. +0.01 points/day), and hopefulness about their health (+0.01 vs. 0.00 points/day). Treatment participants also showed greater increases in perceptions that exercising is easy (+0.1 vs. +0.0 points/week; on a scale from *1: difficult* to *5: easy*).

PA Outcome	Hyp.	Description	Estimate (SE)
Step Count	**H1**	Difference in daily step count from baseline to study period	**+1 680 (307)*****
	H2	Treatment–control difference in daily step count change from baseline to study period	−692 (432)
	H3	Treatment–control difference in the weekly rate of change (Δ steps/day per study week)	+229 (137)
Active Energy Burned (kcal)	**H1**	Difference in daily kcal burned from baseline to study period	**+87.1 (23.4)****
	H2	Treatment–control difference in daily kcal burned change from baseline to study period	−7.09 (31.3)
	H3	Treatment–control difference in the weekly rate of change (Δ kcal/day per study week)	+7.80 (9.82)
Exercise Time (min)	**H1**	Difference in daily exercise min from baseline to study period	**+13.2 (3.76)****
	H2	Treatment–control difference in daily exercise min change from baseline to study period	−2.08 (4.95)
	H3	Treatment–control difference in the weekly rate of change (Δ min/day per study week)	+1.51 (1.57)
Distance Walking/ Running (km)	**H1**	Difference in daily distance walked/run (km) from baseline to study period	**+0.756 (0.140)*****
	H2	Treatment–control difference in daily distance walked/run (km) change from baseline to study period	−0.239 (0.198)
	H3	Treatment–control difference in the weekly rate of change (Δ km/day per study week)	+0.071 (0.063)

Table 7: Wearable Data (Quantitative) Results. We report mean (SE) parameter estimates from our model for each hypothesis. *** denotes $p < 0.001$ and ** denotes $p < 0.01$ (p-values are Holm-adjusted for multiple comparisons). Significant results are **bolded**.

6.2.3 Usability. Usability ratings were high overall, with average System Usability Scale (SUS) scores above 80 (on a 0-100 scale), which corresponds to the 90th percentile.[7] However, scores declined from pre- to post-study in both groups, with the treatment condition exhibiting greater declines (−11.6 vs. −4.7 points; 0-100 scale) and post-study variability (standard deviations of 18.8 vs. 8.8 points). A similar pattern was observed with the Subjective Assessment of Speech System Interfaces (SASSI), a usability measure for speech systems (−0.3 vs. −0.1 points; 5-point Likert scale). These reductions in usability scores could reflect novelty effects and/or usability concerns that arise through continued use. The greater decline in usability in the treatment condition likely reflects inherent challenges in integrating non-deterministic LLM chat with a user interface, which we further discuss in Section 7.4.

6.2.4 User Experience. Participants in the treatment group consistently rated the quality of advice and overall user experience more favorably than the control group at both pre- and post-study measurements. We replicate Jörke et al.'s [55] strong initial impressions of the GPTCoach system (pre-study mean of 4.4 in treatment vs. 4.1 in control; 5-point Likert scale). However, as with usability, ratings declined pre- to post-study in both conditions (−0.4 vs. −0.5 points).

Treatment participants rated shared decision-making more highly at both pre- and post-study measurements, with scores increasing pre/post in treatment while decreasing in control (+0.1 vs. −0.2; 5-point Likert scale), suggesting that goal setting was perceived as a more collaborative activity in the treatment group. The treatment group also reported greater increases in insights into their personal data, measured via the Technology Support Reflection Inventory (TSRI) than the control group (+0.1 vs. −0.3; 5-point Likert scale). In addition, treatment participants also rated their interactions as more human-like (post-study mean of 3.5 vs. 2.83).

6.3 Wearable Data

We now turn to objective measures of PA outcomes, complementing the self-reported survey outcomes. Table 7 presents coefficient means, standard errors, and significance level for each hypothesis from our mixed-effects model. Across all four outcomes—step count, active energy burned (kcal), exercise time (min), and distance walking/running (km)—mean PA during the 4-week study period was significantly higher than mean baseline PA (H1), corresponding to moderate-to-large standardized effects (Cohen's $d = 0.59$-0.75). On average, participants walked 1,680 more steps/day ($p < 0.001$, $d = 0.75$), burned 87.1 more kcal/day ($p < 0.01$, $d = 0.59$), spent 13.2 more minutes/day exercising ($p < 0.01$, $d = 0.60$), and walked 0.76 km/day more ($p < 0.001$, $d = 0.74$) during the study period than at baseline. **These effects translate to substantial increases in activity, doubling the proportion of participants meeting the recommended 150 min/week of exercise** [41], increasing from 36% at baseline to 72% during the study (41% → 71% in control; 31% → 74% in treatment).

We did not find significant treatment–control differences in either the overall daily activity change from baseline to the four-week study period ($p = 0.78$ to 1.0) or the weekly rate of change ($p = 0.77$ to 1.0). However, the results showed consistent directional patterns. For the overall change from baseline to the study period (H2), every outcome had a negative coefficient, indicating that from baseline to the study period, the treatment group experienced somewhat smaller, though not significantly different, average gains than the control group. Regarding weekly persistence (H3), the treatment coefficients are uniformly positive, indicating activity in the treatment condition declined less (or even increased) per study week relative to control.

In summary, H1 was confirmed: both groups showed large and significant increases in PA over the baseline. H2 and H3 were not statistically supported, though consistent directional patterns suggest that the control group achieved slightly larger initial increases, while the treatment group maintained activity levels more consistently across weeks. However, we caution that the present sample was underpowered to detect small differences using statistical testing, and interpretations should be viewed as hypothesis generating rather than confirmatory.

[7] https://www.nngroup.com/articles/measuring-perceived-usability/

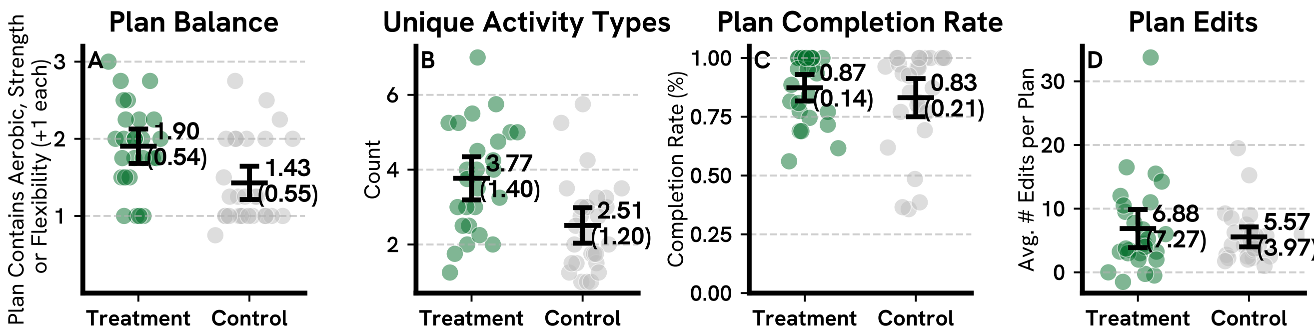

Figure 8: Plan Quality and Personalization by Condition. Participants in the treatment group created more personalized and varied plans compared to control. On average, treatment participants included more activity categories (among aerobic, strength, or flexibility) per plan (1.90 vs. 1.43 categories), a greater number of unique activity types (3.77 vs. 2.51 activity types), and the average plan completion rate was slightly higher in the treatment group (87% vs. 83%). Treatment participants also made more weekly plan edits (6.88 vs. 5.57 edits/week). Error bars represent 95% confidence intervals (1.96 SE).

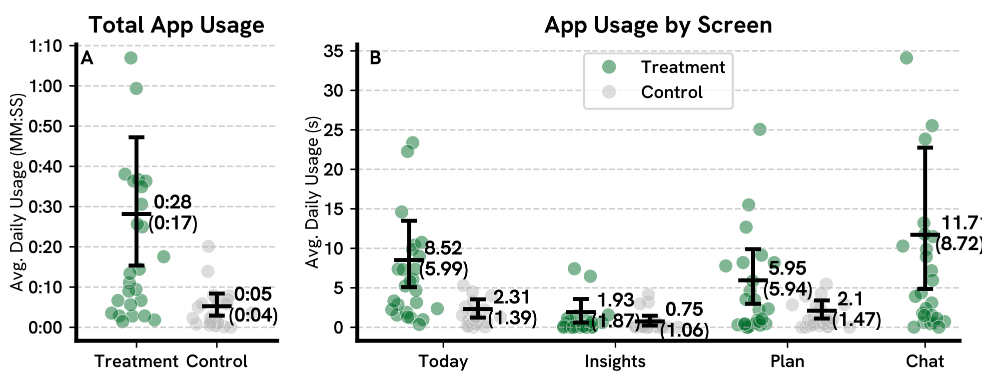

Figure 9: Average Daily Application Usage: Total (A) and By Screen (B). Treatment participants spent over five times more time using the application than control participants. This increased app usage by treatment participants is observed across all screens in the app, indicating that the increase in app use is not merely due to chat-based interaction.

6.4 Plan Data

Next, we examine the quality and personalization of participants' weekly PA plans (Figure 8). First, we assess plan balance by computing a composite score measuring the number of different exercise categories included in each user's weekly plan (+1 for each of aerobic, strength, flexibility; total score out of three), following the American College of Sports Medicine guidelines [97]. Treatment participants had more balanced plans, with a mean (SD) of 1.90 (0.54) categories per weekly plan, compared to 1.43 (0.55) in the control group.

Second, we measure the number of unique activity types per plan. The treatment group exhibited greater variety, with a mean (SD) of 3.77 (1.40) different activity types compared to 2.51 (1.20) in control. Next, we evaluated average plan completion rates (duration of completed activities/duration of all planned activities) and observed a slightly higher completion rate of 87% (14) in the treatment condition compared to control with 83% (21). Finally, we analyze the number of plan edits made by each participant per week. Treatment participants made 6.88 (7.27) plan edits on average, while control made 5.57 (3.97) edits. Together, these results suggest higher plan quality and a higher degree of personalization and engagement in the treatment condition.

6.5 App Usage Data

Finally, we analyze app usage logs to assess engagement. Treatment participants spent 5.6 times as much time in the app as control participants (Figure 9A). This is not only due to LLM chat leading to longer usage times. In Figure 9B, we show that treatment participants showed greater usage times across all screens. This could suggest that users in the treatment condition were more engaged with the app content, possibly due to the more personalized experience enabled by the LLM. As is typical in behavior change interventions, app usage declined over time in both conditions. The treatment group began with higher usage (428 s on day 1) and experienced a steeper daily decrease (-10 s/day), ending at 146 s on day 28, compared to the control group, which began lower (126 s) and declined gradually (-3 s/day), ending at 12 s.

7 Discussion

Our study provides rich insights into how integrating an LLM coaching chatbot with established behavior change interactions can impact PA behaviors and associated psychological outcomes. Bloom's LLM-augmented interventions led to positive changes in participants' mindsets, supported more flexible and personalized planning, and substantially increased app usage duration. In this section, we discuss implications for designing LLM-augmented behavior change interventions, the roles of LLMs in behavioral health systems, and limitations to address in future research.

7.1 The Role of LLMs in Behavioral Health Support: Coaching and Personalization

In Bloom, the LLM served two complementary roles: (1) as a conversational coaching agent that engaged in supportive conversations, and (2) as a personalization mechanism for tailoring interventions based on qualitative context. Whereas most prior work on LLMs in behavioral health has conceptualized their potential primarily in terms of personalization (e.g., customizing plans, optimizing just-in-time nudges, or extracting insights from data), our findings highlight that LLMs' role as a relational agent providing motivational and non-judgmental support was more impactful for participants in early stages of behavior change and low baseline levels of activity.

In our interviews, participants more often attributed their improved mindsets toward PA—increased enjoyment, greater identification with being an active person, greater self-confidence and self-compassion, and nuanced perspectives on what counts as exercise—to Beebo's collaborative and supportive tone. These qualitative insights are consistent with our survey findings, in which treatment participants experienced larger pre-post gains in PA adequacy mindset and satisfaction with PA. Although participants appreciated that the LLM's support was personalized to their unique circumstances, they more often attributed their ability to stick to their weekly plans to the encouragement and flexibility provided by the LLM along the way (e.g., rescheduling missed activities or suggesting alternatives), rather than to the personalized plans themselves. This is mirrored in our quantitative findings, where treatment participants demonstrated greater plan variety and adherence without a corresponding increase in objective PA outcomes.

These findings align with existing behavior change theory, including principles of MI [86] and the TTM [108], both of which emphasize that changing thoughts, feelings, and attitudes toward a behavior is crucial for individuals who are just starting out. Shifts in mindset (i.e., positive beliefs about the benefits of exercise) carry benefits to both psychological and physiological health [34] independent of immediate behavior change. Moreover, these changes in mindset may also represent shifts in decisional balance (i.e., stronger perceived pros and fewer perceived cons of PA), which is predictive of sustainable, longer-term behavior change in the TTM [108]. Thus, while immediate quantitative outcomes were not significantly different, the impact of LLM coaching on mindsets could potentially yield measurable behavioral changes over an extended timeframe, warranting longer-term studies.

However, it remains an open question whether more granular LLM-driven personalization or more prescriptive coaching agents might better serve individuals who are already highly active or in later stages of change. These individuals may require more specific, performance-oriented support. Nonetheless, populations in early stages of change and/or with low baseline activity levels are particularly important to support as they stand to benefit most from behavioral health interventions.

7.2 Leveraging Qualitative Context to Foster Agency

While participants rarely identified Beebo as the single most useful feature, they described its role in supporting and reinforcing accountability provided by other app features, such as plans and notifications. Across our interviews, participants consistently valued the way Beebo's interactions helped them feel a stronger sense of *agency* over their PA: agency in setting their own goals, adjusting plans as barriers arose, and reframing everyday activities as legitimate forms of exercise.

Participants highlighted Beebo's role in collaborative goal-setting and appreciated how Beebo asked to reschedule missed activities rather than abandon them, or offered to substitute alternatives that better matched their schedules or abilities. Participants enjoyed notifications from Beebo that encouraged reflection on progress and motivation, which were perceived not as tasks to be completed but reminders of their own goals and motivations. Survey results bolster these findings, as treatment participants reported greater shared decision-making around goal-setting and perceived greater insight into their personal data.

Many of the interactions participants valued most were only possible because Beebo could leverage qualitative context expressed in natural language. For example, collaborative goal-setting requires that the agent asks about and adapts to personal constraints (e.g., childcare or social commitments) or physical abilities. Notifications that remind participants of their long-term goals and motivations necessitate knowledge of these goals and motivations. Others described how Beebo helped reframe activities they already enjoyed (e.g., gardening or walking at the mall) as legitimate forms of exercise, which reduced the perceived burden of "exercising" and made PA feel more achievable.

These findings highlight a design opportunity for LLM-augmented interactions to enhance agency not simply by personalizing content, but by contextualizing support in ways that help participants recognize and build upon their own capacities. Promoting agency is also strongly aligned with principles of health coaching and MI. For future systems, this could include integrating additional context sources (e.g., calendar or weather data) to proactively anticipate barriers, or reframing data visualizations around participants' own goals to celebrate progress on their own terms rather than merely reporting quantitative metrics. Our results suggest that LLM augmentation is particularly useful for agency-enhancing interventions because qualitative context allows interventions to feel personally relevant and aligned with participants' own motivations.

7.3 Social and Relational Cues: Opportunities and Risks for Promoting Engagement

Participants in the treatment condition exhibited increased engagement across all screens of the Bloom app. While multiple factors may have contributed to this, interview feedback suggests that the perceived social and relational nature of interactions with Beebo played an important role. Many participants explicitly said that chatting with Beebo felt like talking to a person, even while recognizing it was an AI. This mirrors findings from pre-LLM health coaching systems [15, 87], as well as recent LLM-based coaching [55], and aligns with long-standing evidence that people respond socially to computers and media [109]. Participants identified several qualities that fostered this sense of presence, mentioning the bee avatar, Beebo's empathetic and non-judgmental tone, proactive but optional messaging, and references back to prior conversations. Unlike the control condition, LLM-generated notifications and chatbot interactions were frequently described as interactions with an accountability partner rather than simple reminders.

These social and relational cues carry both benefits and risks. On the one hand, they helped participants maintain accountability and fostered positive mindsets around exercise. One implication of this is that increasing anthropomorphic qualities by using more personalized and empathetic messaging, human-like avatars, or expanding conversational topics beyond strictly health-related content could strengthen engagement and accountability. On the other hand, such designs may foster emotional attachment or over-reliance on the chatbot, raising concerns about self-efficacy and intrinsic motivation. Prior work cautions that support from chatbots can differ in important and controversial ways from support provided by humans [103] and recent studies point to the mental health risks associated with affective use of LLM chatbots [80, 106]. Conversely, interfaces that minimize anthropomorphism or rely solely on UI-based, non-conversational interactions (e.g., simple free-response input) may mitigate these risks but may lose out on the motivational and accountability benefits we observed.

Importantly, our findings show that social and relational cues should not be seen as the only means to foster engagement. For instance, Bloom's ambient display was perceived as playful and appealing, was highly popular, and contributed to engagement and motivation. Prior work on narrative for behavior change [93] suggests that incorporating narrative structures, where an LLM avatar could serve as a character in a story or interactive guide, could further enhance engagement and motivation without relying on increased anthropomorphism. Additionally, integrating LLM-augmented interactions within systems that facilitate human social support, where LLMs might mediate or support peer-to-peer interactions, offers another promising approach to foster accountability and engagement [87].

Overall, we do not advocate for replicating or replacing human connection and social support with LLMs. Instead, we argue that the benefits of relational cues in LLM coaching must be weighed against their risks, recognizing that automated and human health coaching are distinct and entail different affordances.

7.4 Design Challenges and Future Directions

While Bloom received largely positive feedback, participants identified several areas for improvement. For example, Beebo's verbosity and repetitiveness were common complaints. In preliminary experiments, shortening responses via prompting resulted in reduced perceived empathy and support. We decided to allow our agent to produce longer responses based on this pilot feedback, even though verbosity is associated with lower-quality counseling [105]. Addressing verbosity concerns will likely require larger scale data collection and fine-tuning beyond the prompting approaches used in this study. Meanwhile, our notification generation prompt included the last ten notifications along with a diversity prompt, yet they were still perceived as repetitive by some participants. Future iterations could incorporate more sophisticated memory systems to minimize repetitive or overly generic responses.

Similarly, interactions sometimes felt rigid or "bot-like," potentially stemming from our dialogue state chain enforcing hard bounds on the conversation topic. While removing the dialogue state chain allowed the agent to more naturally adapt to a conversation, it also allowed it to veer off course and lose track of the conversation's goals, mirroring prior findings in multi-turn settings [69]. Future research on goal-directed, multi-turn conversations could improve LLMs' conversational flexibility while ensuring they achieve conversational goals and stay within fixed conversational bounds.

In addition, we encountered several issues attributable to inconsistent tool use. Participants occasionally experienced frustrating interactions due to hallucinated or inaccurate tool calls, particularly around scheduling and activity tracking. For example, hallucinated tool calls (e.g., extraneous or incorrect arguments) could lead to errors in plan edits, upon which the agent would tell the user they would "fix the issue later," even though it did not have this capability. Moreover, health data fetching was rarely proactively initiated by the model, mirroring findings from GPTCoach [55]. Improving the robustness of tool use will likely also necessitate fine-tuning.

Finally, several usability challenges emerged from inherent challenges in integrating conversational agents with complex user interfaces. The open-ended chat interface and Beebo's conversational flexibility surfaced several false affordances: participants frequently posed questions about app features or bugs directly to Beebo (e.g., why workouts were not syncing, summaries not loading, and check-in not starting), which Beebo frequently did not have the knowledge or capabilities to address. Adding to their frustration, the agent would sometimes hallucinate an incorrect answer instead of stating that it did not know. These results might also explain the larger variance in post-study usability scores in the treatment condition. Addressing these concerns requires both clear communication of agent capabilities, either in the UI or chat, to calibrate participants' mental models as well as preventing problematic model hallucinations.

Taken together, while our findings motivate a longer-term (several months to a year), larger-scale (hundreds of participants) study to evaluate whether larger behavioral differences materialize over time, several targeted, currently feasible improvements to the LLM agent are likely to increase the effect size.

8 Limitations

Our study has several limitations. First, our study design was not intended to robustly evaluate behavior change in a clinical sense, which requires large-scale, randomized controlled trials conducted over several months to years. Consequently, our statistical tests were likely underpowered, and our study duration was likely insufficient for larger differences in PA habits to materialize. Since participants self-selected into the study and were compensated, the effect sizes observed here are likely larger than what would be expected in a naturalistic deployment. We concur with Klasnja et al. [63], who have argued that evaluating behavior change in a traditional sense is often inappropriate for early-stage HCI systems research. Instead, our study's main focus was on participant experiences through qualitative interviews and extensive surveys, aiming to uncover insights into why and how such systems succeed or fail. The four-week duration balanced high participation demands with known issues with sustained engagement in behavior change research [140]. Further, our sample size is larger than is typical in previous HCI deployment studies targeting PA promotion [32, 91] or LLM health coaching [138].

Since our control condition did not include a rule-based chatbot, some of our findings may have also arisen with a non-LLM chatbot. However, our findings around relational cues and mindsets differ from prior work on rule-based coaches [76, 87]. Moreover, given that we did not observe significant differences in quantitative PA outcomes with an LLM chatbot, it is unlikely that a simpler chatbot would have produced stronger effects.

Our recruitment procedure aimed for broad demographic coverage, requiring over 2,000 screening responses to enroll 54 participants. Although our sample had adequate diversity along gender, age, and race/ethnicity, participants had higher educational attainment and income levels compared to the general population. This likely reflects our requirements for owning an iPhone and Apple Watch [54] as well as the significant time commitment required for participation [66]. While automated health coaching has the potential to particularly benefit individuals experiencing greater barriers to PA, often linked to lower socioeconomic status (SES) [131], future research that specifically engages with low-SES populations is necessary to realize this potential and prevent intervention-generated inequalities [130]. This may involve community-based recruitment procedures and/or relaxing device ownership requirements.

Lastly, the Bloom system itself had technological limitations. We did not explore more advanced agent designs involving finetuning [57, 77], multi-agent systems [81], or reasoning models [47, 53]. While our summary-based memory module was well within the model's context limit, even for users that chatted daily (<3,500 tokens), more sophisticated memory systems (e.g., [104]) could additionally have improved performance. Given the promising initial feedback we received on our prompt-based approach, we considered it sufficient for generating design insights. Moreover, advanced improvements would have required significant development and/or data collection efforts without clear evidence of user need or interaction requirements. We also did not explore all possible LLM-augmented behavior change interactions, such as social features or gamification. Nevertheless, we believe that qualitative context can be informative for a broad range of interactions and is a valuable exploration for future work.

9 Conclusion

In this work, we introduce Bloom, a mobile application integrating an LLM-based coaching chatbot with established behavior change interactions. Bloom leverages qualitative context from coaching conversations to further personalize LLM-augmented behavior change interactions, moving beyond existing text-only approaches in prior work on LLMs for behavioral health. Through a four-week field study involving 54 participants, we found that Bloom fostered positive mindsets toward PA and supported more personalized and flexible workout planning compared to a non-LLM control. Treatment participants reported greater improvements in psychological and motivational outcomes, including increased enjoyment, self-confidence, and a greater sense of agency in managing their PA goals. Both conditions substantially increased participants' exercise levels, doubling the number of individuals who met or exceeded the recommended 150 min/week of PA. Although we observed no significant differences between conditions in objectively measured PA, psychological and motivational outcomes indicate the potential for sustained behavior change over longer periods when using our LLM-based health coach. Bloom presents a novel opportunity to create more supportive and empowering behavior change applications that leverage qualitative context to more sustainably shape people's mindsets and motivation.

10 Acknowledgments

We are grateful for the funding support provided by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and the Hasso Plattner Foundation, as well as the OpenAI researcher access program for providing API credits to partially support this research. We thank the many members of the IxD and AI4HI research groups for their continuous support and feedback on this project, and all of our friends & family testers for catching bugs and suggesting improvements. We thank Michelle Lam, Lindsay Popowski, and Dulce Garcia for their assistance with participant recruitment; Ryan Louie, Ifdita Hasan, and Ankita Koodavoor for their advice on our redteaming and safety evaluation; Andrew Sung for his contributions to app development; and Carolyn Zhou for their feedback on the LLM agent architecture. We are also grateful to Tobias Gerstenberg, Nilam Ram, and Michael Bernstein for guidance on statistical modeling, to Elizabeth Murnane for her input on our ambient display design and validation study, and to Mary Czerwinski, Andrea Green, Omar Shaikh, and Helena Vasconcelos for thoughtful feedback on early drafts of this manuscript. Most importantly, we thank all of our participants, without whom this research would not have been possible.

References

[1] Elena Agapie, Bonnie Chinh, Laura R Pina, Diana Oviedo, Molly C Welsh, Gary Hsieh, and Sean Munson. 2018. Crowdsourcing exercise plans aligned with expert guidelines and everyday constraints. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Tim Althoff, Ryen W White, and Eric Horvitz. 2016. Influence of Pokémon Go on physical activity: study and implications. *Journal of Medical Internet Research* 18, 12 (2016), e315.

[3] Ian Anderson, Julie Maitland, Scott Sherwood, Louise Barkhuus, Matthew Chalmers, Malcolm Hall, Barry Brown, and Henk Muller. 2007. Shakra: tracking and sharing daily activity levels with unaugmented mobile phones. *Mobile networks and applications* 12 (2007), 185–199.

[4] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. 2020. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–9.

[5] Michelle Bak and Jessie Chin. 2024. The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. *Journal of the American Medical Informatics Association* 31, 9 (2024), 2047–2053.

[6] Albert Bandura. 1999. Social cognitive theory: An agentic perspective. *Asian journal of social psychology* 2, 1 (1999), 21–41.

[7] Eric PS Baumer. 2015. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 585–594.

[8] Erkan Başar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A Bosch, and Emiel Krahmer. 2025. How well can large language models reflect? A human evaluation of LLM-generated reflections for motivational interviewing dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*. 1964–1982.

[9] Frank Bentley and Konrad Tollmar. 2013. The power of mobile notifications to increase wellbeing logging behavior. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1095–1098.

[10] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 1–27.

[11] Marit Bentvelzen, Jasmin Niess, Mikołaj P Woźniak, and Paweł W Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[12] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[13] Timothy Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics* 39, 5 (2006), 556–571.

[14] Timothy W Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in computing systems*. 1212–1215.

[15] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 293–327.

[16] Danielle Z Boles, Maysa DeSousa, Bradley P Turnwald, Rina I Horii, Taylor Duarte, Octavia H Zahrt, Hazel R Markus, and Alia J Crum. 2021. Can exercising and eating healthy be fun and indulgent instead of boring and depriving? Targeting mindsets about the process of engaging in healthy behaviors. *Frontiers in psychology* 12 (2021), 745950.

[17] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]

[18] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[19] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[20] Lorraine R Buis, Timothy A Poulton, Robert G Holleman, Ananda Sen, Paul J Resnick, David E Goodrich, LaVaughn Palma-Davis, and Caroline R Richardson. 2009. Evaluating Active U: an Internet-mediated physical activity program. *BMC public health* 9, 1 (2009), 331.

[21] Fiona C Bull, Salih S Al-Ansari, Stuart Biddle, Katja Borodulin, Matthew P Buman, Greet Cardon, Catherine Carty, Jean-Philippe Chaput, Sebastien Chastin, Roger Chou, et al. 2020. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British journal of sports medicine* 54, 24 (2020), 1451–1462.

[22] Cynthia M Castro, Leslie A Pruitt, Matthew P Buman, and Abby C King. 2011. Physical activity program delivery by professionals versus volunteers: the TEAM randomized trial. *Health Psychology* 30, 3 (2011), 285.

[23] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *arXiv preprint arXiv:2401.00820* (2024).

[24] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. 2022. Reflection in theory and reflection in practice: An exploration of the gaps in reflection support among personal informatics apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[25] 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.

[26] Eun Kyoung Choe, Bongshin Lee, et al. 2015. Characterizing visualization insights from quantified selfers' personal data presentations. *IEEE computer graphics and applications* 35, 4 (2015), 28–37.

[27] Akshat Choube, Ha Le, Jiachen Li, Kaixin Ji, Vedant Das Swain, and Varun Mishra. 2025. GLOSS: Group of LLMs for Open-Ended Sensemaking of Passive Sensing Data for Health and Wellbeing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–32.

[28] Sunny Consolvo. 2008. *Designing and evaluating a persuasive technology to encourage lifestyle behavior change*. PhD Dissertation. University of Washington, Seattle, WA.

[29] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A Landay. 2006. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 457–466.

[30] Sunny Consolvo, Predrag Klasnja, David W McDonald, James A Landay, et al. 2014. Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Foundations and Trends® in Human–Computer Interaction* 6, 3–4 (2014), 167–315.

[31] Sunny Consolvo, David W McDonald, and James A Landay. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 405–414.

[32] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1797–1806.

[33] Cora L Craig, Alison L Marshall, Michael Sjöström, Adrian E Bauman, Michael L Booth, Barbara E Ainsworth, Michael Pratt, ULF Ekelund, Agneta Yngve, James F Sallis, et al. 2003. International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise* 35, 8 (2003), 1381–1395.

[34] Alia J Crum and Ellen J Langer. 2007. Mind-set matters: Exercise and the placebo effect. *Psychological science* 18, 2 (2007), 165–171.

[35] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons learned from two cohorts of personal informatics self-experiments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–22.

[36] Tina Ekhtiar, Armağan Karahanoğlu, Rúben Gouveia, and Geke Ludden. 2023. Goals for goal setting: a scoping review on personal informatics. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2625–2641.

[37] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Chun-Cheng Chang, Xuhai" Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.

[38] Daniel A Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, et al. 2020. Mapping and taking stock of the personal informatics literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–38.

[39] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. Physiollm: Supporting

personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 1–8.

[40] Mohammad Feli, Iman Azimi, Pasi Liljeberg, and Amir M Rahmani. 2025. An LLM-Powered Agent for Physiological Data Analysis: A Case Study on PPG-based Heart Rate Estimation. *arXiv preprint arXiv:2502.12836* (2025).

[41] Centers for Disease Control and Prevention. 2022. Physical Activity. https://www.cdc.gov/physicalactivity/index.html

[42] Roland Gasser, Dominique Brodbeck, Markus Degen, Jürg Luthiger, Remo Wyss, and Serge Reichlin. 2006. Persuasiveness of a mobile lifestyle coaching application using social facilitation. In *Persuasive Technology: First International Conference on Persuasive Technology for Human Well-Being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 18-19, 2006. Proceedings 1*. Springer, 27–38.

[43] Stefan Göbel, Sandro Hardy, Viktor Wendel, Florian Mehm, and Ralf Steinmetz. 2010. Serious games for health: personalized exergames. In *Proceedings of the 18th ACM international conference on Multimedia*. 1663–1666.

[44] Peter M Gollwitzer. 1999. Implementation intentions: strong effects of simple plans. *American psychologist* 54, 7 (1999), 493.

[45] Peter M Gollwitzer and Paschal Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology* 38 (2006), 69–119.

[46] Rúben Gouveia, Fábio Pereira, Evangelos Karapanos, Sean A Munson, and Marc Hassenzahl. 2016. Exploring the design space of glanceable feedback for physical activity trackers. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 144–155.

[47] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[48] A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, et al. 2025. The Anatomy of a Personal Health Agent. *arXiv preprint arXiv:2508.20148* (2025).

[49] Jennifer L Hicks, Melissa A Boswell, Tim Althoff, Alia J Crum, Joy P Ku, James A Landay, Paula ML Moya, Elizabeth L Murnane, Michael P Snyder, Abby C King, et al. 2023. Leveraging mobile technology for public health promotion: A multidisciplinary perspective. *Annual Review of Public Health* 44 (2023), 131–150.

[50] Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (2000), 287–303.

[51] Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2024. Llasa: A multimodal llm for human activity analysis through wearable and smartphone sensors. *arXiv preprint arXiv:2406.14498* (2024).

[52] Stephen S Intille. 2004. Ubiquitous computing technology for just-in-time motivation of behavior change. In *MEDINFO 2004*. IOS Press, 1434–1437.

[53] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

[54] Maral Jamalova. 2019. The comparative study of the relationship between smartphone choice and socio-economic indicators. *Int. J. Mark. Stud* 11, 11 (2019), 10–5539.

[55] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–46.

[56] Elisabeth T Kersten-van Dijk, Joyce HDM Westerink, Femke Beute, and Wijnand A IJsselsteijn. 2017. Personal informatics, self-insight, and behavior change: A critical review of current literature. *Human–Computer Interaction* 32, 5-6 (2017), 268–296.

[57] Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. 2025. A personal health large language model for sleep and fitness coaching. *Nature Medicine* (2025), 1–10.

[58] Abby C King, Maria Ines Campero, Jylana L Sheats, Cynthia M Castro Sweet, Michelle E Hauser, Dulce Garcia, Aldo Chazaro, German Blanco, Jorge Banda, David K Ahn, et al. 2020. Effects of counseling by peer human advisors vs computers to increase walking in underserved populations: The COMPASS randomized clinical trial. *JAMA internal medicine* 180, 11 (2020), 1481–1490.

[59] Abby C King, Robert Friedman, Bess Marcus, Cynthia Castro, Melissa Napolitano, David Ahn, and Lawrence Baker. 2007. Ongoing physical activity advice by humans versus computers: the Community Health Advice by Telephone (CHAT) trial. *Health Psychology* 26, 6 (2007), 718.

[60] Abby C. King, William L. Haskell, C. Barr Taylor, Robert DeBusk, Cynthia M. Castro, Leslie A. Pruitt, and Stanford Prevention Research Center staff. 2002. The Stanford Active Choices Program: Telephone-Assisted Counseling for Physical Activity. Stanford Health Promotion Resource Center, Stanford Prevention Research Center, Stanford University.

[61] Abby C King, Eric B Hekler, Cynthia M Castro, Matthew P Buman, Bess H Marcus, Robert H Friedman, and Melissa A Napolitano. 2014. Exercise advice by humans versus computers: maintenance effects at 18 months. *Health Psychology* 33, 2 (2014), 192.

[62] Predrag Klasnja, Sunny Consolvo, David W McDonald, James A Landay, and Wanda Pratt. 2009. Using mobile & personal sensing technologies to support health behavior change in everyday life: lessons learned. In *AMIA Annual Symposium Proceedings*, Vol. 2009. 338.

[63] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3063–3072. https://doi.org/10.1145/1978942.1979396

[64] Predrag Klasnja and Wanda Pratt. 2012. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of biomedical informatics* 45, 1 (2012), 184–198.

[65] William J Korotitsch and Rosemery O Nelson-Gray. 1999. An overview of self-monitoring research in assessment and treatment. *Psychological Assessment* 11, 4 (1999), 415.

[66] Whitney-Jocelyn Kouaho and Daniel A Epstein. 2024. Socioeconomic Class in Physical Activity Wearables Research and Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[67] Levente Kriston, Isabelle Scholl, Lars Hölzel, Daniela Simon, Andreas Loh, and Martin Härter. 2010. The 9-item Shared Decision Making Questionnaire (SDM-Q-9). Development and psychometric properties in a primary care sample. *Patient education and counseling* 80, 1 (2010), 94–99.

[68] Ash Tanuj Kumar, Cindy Wang, Alec Dong, and Jonathan Rose. 2024. Generation of Backward-Looking Complex Reflections for a Motivational Interviewing–Based Smoking Cessation Chatbot Using GPT-4: Algorithm Development and Validation. *JMIR Mental Health* 11, 1 (2024), e53778.

[69] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120* (2025).

[70] Nicholas Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2012. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*.

[71] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.

[72] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Context-Driven Sensemaking of Multi-modal Personal Tracking Data Using Visualization and Human-in-the-Loop LLM. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–37.

[73] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings 8*. Springer, 261–278.

[74] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. arXiv:2305.15525 [cs.CL]

[75] Edwin A Locke and Gary P Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist* 57, 9 (2002), 705.

[76] Tiffany Christina Luo, Adrian Aguilera, Courtney Rees Lyles, and Caroline Astrid Figueroa. 2021. Promoting physical activity through conversational agents: mixed methods systematic review. *Journal of Medical Internet Research* 23, 9 (2021), e25486.

[77] Sriya Mantena, Anders Johnson, Marily Oppezzo, Narayan Schuetz, Alexander Tolas, Ritu Doijad, C Mikael Mattson, Allan Lawrie, Mariana Ramirez-Posada, Eleni Linos, et al. 2025. Fine-tuning Large Language Models in Behavioral Psychology for Scalable Physical Activity Coaching. *medRxiv* (2025).

[78] Bess H Marcus, Vanessa C Selby, Raymond S Niaura, and Joseph S Rossi. 1992. Self-efficacy and the stages of exercise behavior change. *Research quarterly for exercise and sport* 63, 1 (1992), 60–66.

[79] Tara Lynn Matthews. 2007. Designing and Evaluating Glanceable Peripheral Displays. (2007).

[80] Miles McCain, Ryn Linthicum, Chloe Lubinski, Alex Tamkin, Saffron Huang, Michael Stern, Kunal Handa, Esin Durmus, Tyler Neylon, Stuart Ritchie, Kamya Jagadish, Paruul Maheshwary, Sarah Heck, Alexandra Sanderford, and Deep Ganguli. 2025. *How People Use Claude for Support, Advice, and Companionship*. https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship

[81] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al.

2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464* (2024).

[82] Selina Meyer and David Elsweiler. 2025. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies* 200 (2025), 103514.

[83] Sophia Meywirth, Andreas Janson, and Matthias Söllner. 2025. Designing for Trust: Integrating Self-referencing in Large Language Model-Based Health Coaching. In *International Conference on Design Science Research in Information Systems and Technology*. Springer, 296–309.

[84] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine* 46, 1 (2013), 81–95.

[85] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6, 1 (2011), 42.

[86] W.R. Miller and S. Rollnick. 2023. *Motivational Interviewing: Helping People Change and Grow* (4th ed.). Guilford Publications.

[87] Elliot G Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N Tobin, Patricia Davidson, Arlene M Smaldone, and Lena Mamykina. 2021. Automated vs. human health coaching: exploring participant and practitioner experiences. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–37.

[88] David C Mohr, Stephen M Schueller, Enid Montague, Michelle Nicole Burns, and Parisa Rashidi. 2014. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. *Journal of Medical Internet Research* 16, 6 (2014), e146.

[89] Theresa Moyers, Tim Martin, Delwyn Catley, Kari Jo Harris, and Jasjit S Ahluwalia. 2003. Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy* 31, 2 (2003), 177–184.

[90] Sean A Munson and Sunny Consolvo. 2012. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In *2012 6th international conference on pervasive computing technologies for healthcare (pervasivehealth) and workshops*. IEEE, 25–32.

[91] Elizabeth L Murnane, Yekaterina S Glazko, Jean Costa, Raymond Yao, Grace Zhao, Paula ML Moya, and James A Landay. 2023. Narrative-based visual feedback to encourage sustained physical activity: A field trial of the WhoIsZuki mobile health platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–36.

[92] Elizabeth L Murnane, David Huffaker, and Gueorgi Kossinets. 2015. Mobile health apps: adoption, adherence, and abandonment. In *Adjunct proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 ACM international symposium on wearable computers*. 261–264.

[93] Elizabeth L Murnane, Xin Jiang, Anna Kong, Michelle Park, Weili Shi, Connor Soohoo, Luke Vink, Iris Xia, Xin Yu, John Yang-Sammataro, et al. 2020. Designing ambient narrative-based interfaces to reflect and motivate physical activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[94] 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* (2018), 1–17.

[95] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[96] Cameron D Norman and Harvey A Skinner. 2006. eHEALS: the eHealth literacy scale. *Journal of Medical Internet Research* 8, 4 (2006), e507.

[97] American College of Sports Medicine. 2025. *ACSM's Guidelines for Exercise Testing and Prescription* (12 ed.). Wolters Kluwer Health.

[98] Jeanette M Olsen. 2014. Health coaching: a concept analysis. In *Nursing forum*, Vol. 49. Wiley Online Library, 18–29.

[99] Jeanette M Olsen and Bonnie J Nesbitt. 2010. Health coaching to improve healthy lifestyle behaviors: an integrative review. *American journal of health promotion* 25, 1 (2010), e1–e12.

[100] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[101] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[102] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*. 129–136.

[103] Shuyi Pan and Maartje MA De Graaf. 2025. Developing a Social Support Framework: Understanding the Reciprocity in Human-Chatbot Relationship. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–13.

[104] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[105] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 926–935.

[106] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).

[107] Katrina L Piercy, Richard P Troiano, Rachel M Ballard, Susan A Carlson, Janet E Fulton, Deborah A Galuska, Stephanie M George, and Richard D Olson. 2018. The physical activity guidelines for Americans. *Jama* 320, 19 (2018), 2020–2028.

[108] James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion* 12, 1 (1997), 38–48.

[109] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996).

[110] Betsy Rodríguez, Jane Kelly, and Michelle Owens. 2008. Road to health toolkit user's guide. (2008).

[111] Jonathan Rose, Imtihan Ahmed, Eric Keilty, Carolynne Cooper, and Peter Selby. 2022. Generation and classification of motivational-interviewing-style reflections for smoking behaviour change using few-shot learning with transformers. *Authorea Preprints* (2022).

[112] James F Sallis, Robin B Pinski, Robin M Grossman, Thomas L Patterson, and Philip R Nader. 1988. The development of self-efficacy scales for healthrelated diet and exercise behaviors. *Health education research* 3, 3 (1988), 283–292.

[113] Paul Schmiedmayer, Vishnu Ravi, and Oliver Aalami. 2024. *Spezi*. https://doi.org/10.5281/zenodo.10482368

[114] Yasaman S Sefidgar, Matthew Jörke, Jina Suh, Koustuv Saha, Shamsi Iqbal, Gonzalo Ramos, and Mary Czerwinski. 2024. Improving work-nonwork balance with data-driven implementation intention and mental contrasting. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (2024), 1–29.

[115] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.

[116] Ali Shameli, Tim Althoff, Amin Saberi, and Jure Leskovec. 2017. How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In *Proceedings of the 26th international conference on world wide web companion*. 455–463.

[117] Jeff Sinclair, Philip Hingston, and Martin Masek. 2007. Considerations for the design of exergames. In *Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia*. 289–295.

[118] Ben Singh, Timothy Olds, Jacinta Brinsley, Dot Dumuid, Rosa Virgara, Lisa Matricciani, Amanda Watson, Kimberley Szeto, Emily Eglitis, Aaron Miatke, et al. 2023. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *npj Digital Medicine* 6, 1 (2023), 118.

[119] Haochen Song, Dominik Hofer, Rania Islambouli, Laura Hawkins, Ananya Bhattacharjee, Meredith Franklin, and Joseph Jay Williams. 2025. Investigating the Relationship Between Physical Activity and Tailored Behavior Change Messaging: Connecting Contextual Bandit with Large Language Models. *arXiv preprint arXiv:2506.07275* (2025).

[120] Vidya Srinivas, Xuhai Xu, Xin Liu, Kumar Ayush, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. 2025. Substance over style: Evaluating proactive conversational coaching agents. *arXiv preprint arXiv:2503.19328* (2025).

[121] Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–10.

[122] Stefan Stepanovic and Tobias Mettler. 2018. Gamification applied for health promotion: does it really foster long-term engagement? A scoping review. In *Proceedings of the 26th European Conference on Information Systems*. AIS, 1–16.

[123] Ajan Subramanian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. 2024. Graph-augmented llms for personalized health insights: A case study in sleep analysis. In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*. IEEE, 1–4.

[124] Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A Bosch. 2024. Rethinking the alignment of psychotherapy

dialogue generation with motivational interviewing strategies. *arXiv preprint arXiv:2408.06527* (2024).

[125] Alice Thudt, Dominikus Baur, Samuel Huron, and Sheelagh Carpendale. 2015. Visual mementos: Reflecting memories with personal data. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 369–378.

[126] Tammy Toscos, Anne Faber, Shunying An, and Mona Praful Gandhi. 2006. Chick clique: persuasive technology to motivate teenage girls to exercise. In *CHI'06 extended abstracts on Human factors in computing systems*. 1873–1878.

[127] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters* 36 (2014), 189–195.

[128] Corneel Vandelanotte, Stewart Trost, Danya Hodgetts, Tasadduq Imam, Mamunur Rashid, Quyen G To, and Carol Maher. 2023. Increasing physical activity using an just-in-time adaptive digital assistant supported by machine learning: a novel approach for hyper-personalised mHealth interventions. *Journal of Biomedical Informatics* 144 (2023), 104435.

[129] Madhurima Vardhan, Narayan Hegde, Deepak Nathani, Emily Rosenzweig, Cathy Speed, Alan Karthikesalingam, and Martin Seneviratne. 2025. Correction: Infusing behavior science into large language models for activity coaching. *PLOS Digital Health* 4, 3 (2025), e0000786.

[130] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.

[131] Adam Wagstaff. 2002. Poverty and health sector inequalities. *Bulletin of the world health organization* 80 (2002), 97–105.

[132] Xingbo Wang, Janessa Griffith, Daniel A Adler, Joey Castillo, Tanzeem Choudhury, and Fei Wang. 2025. Exploring Personalized Health Support through Data-Driven, Theory-Guided LLMs: A Case Study in Sleep Health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[133] Thomas Webb, Judith Joseph, Lucy Yardley, Susan Michie, et al. 2010. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of Medical Internet Research* 12, 1 (2010), e1376.

[134] Sara Wilcox, Marsha Dowda, Sarah F Griffin, Carol Rheaume, Marcia G Ory, Laura Leviton, Abby C King, Andrea Dunn, David M Buchner, Terry Bazzarre, et al. 2006. Results of the first year of active for life: translation of 2 evidence-based physical activity programs for older adults into community settings. *American Journal of Public Health* 96, 7 (2006), 1201–1209.

[135] Sara Wilcox, Marsha Dowda, Laura C Leviton, Jenny Bartlett-Prescott, Terry Bazzarre, Kimberly Campbell-Voytal, Ruth Ann Carpenter, Cynthia M Castro, Diane Dowdy, Andrea L Dunn, et al. 2008. Active for life: final results from the translation of two physical activity programs. *American journal of preventive medicine* 35, 4 (2008), 340–351.

[136] Ruth Q Wolever, Leigh Ann Simmons, Gary A Sforzo, Diana Dill, Miranda Kaye, Elizabeth M Bechard, Mary Elaine Southard, Mary Kennedy, Justine Vosloo, and Nancy Yang. 2013. A systematic review of the literature on health and wellness coaching: defining a key behavioral intervention in healthcare. *Global advances in health and medicine* 2, 4 (2013), 38–57.

[137] Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot Dialogue Strategy Learning for Motivational Interviewing via Inductive Reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*. 13207–13219.

[138] Yan Xu, Brennan Jones, Hannah Nguyen, Qisheng Li, and Stefan Scherer. 2025. From Goals to Actions: Designing Context-aware LLM Chatbots for New Year's Resolutions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1–17.

[139] Qinghua Yang and Stephanie K Van Stee. 2019. The comparative effectiveness of mobile phone interventions in improving health outcomes: meta-analytic review. *JMIR mHealth and uHealth* 7, 4 (2019), e11244.

[140] Xiaotian Yang, Lin Ma, Xi Zhao, and Atreyi Kankanhalli. 2020. Factors influencing user's adherence to physical activity applications: A scoping literature review and future directions. *International Journal of Medical Informatics* 134 (2020), 104039.

[141] Octavia H Zahrt and Alia J Crum. 2020. Effects of physical activity recommendations on mindset, behavior and perceived health. *Preventive medicine reports* 17 (2020), 101027.

[142] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. 2025. SensorLM: Learning the Language of Wearable Sensors. *arXiv preprint arXiv:2506.09108* (2025).

[143] Xinzhe Zheng, Sijie Ji, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. 2025. ProMind-LLM: Proactive Mental Health Care via Causal Reasoning with Sensor Data. *arXiv preprint arXiv:2505.14038* (2025).

A LLM Coaching Agent

In this section, we provide further technical and implementation details on our LLM agent's tools.

A.1 Querying Health Data

The `query_health_data` function allows the agent to fetch the user's wearable data leveraging Apple's HealthKit API. This function is always available to the agent and has the following function signature:

```
query_health_data(
  sample_type: string,
  reference_date: string = ‘today’,
  aggregation_level: enum[‘day’|‘week’|‘month’] = ‘month’,
  show_user: boolean = false
)
```

The HealthKit query is executed on-device in Swift and returns the aggregated data as text along with a natural language description of the data source to the LLM. If `show_user=true`, the data is displayed as a visualization in the chat (see Figure 3D). We encountered issues with LLMs being able to reliably perform date arithmetic, so we allow `reference_date` to be parsed either as a natural language string (e.g., "today", "last week") or a date string. Unlike GPTCoach [55], we do not include a tool use chain (a prompt chain that forces the model to decide whether to augment its response with the user's health data), which led to many unnecessary data queries and greatly increased latency.

A.2 Structured Plan Generation

During onboarding and check-in chats, we expose the `generate_plan` function to the agent. This function generates a weekly plan using an external prompt that takes the current conversation history and the user's plan history as input and produces a structured JSON object. After generating the plan as a JSON object, it is displayed to the user as a plan widget in the chat (see Figure 2D; Figure 3D) and saved to our database such that it can be rendered and modified in the app's UI.

The plan generation prompt includes guidelines for creating a well-rounded, stage-appropriate, and personalized exercise plans sourced from the Active Choice program and the Center for Disease Control's Physical Activity Guidelines [41]. The agent is instructed to call the `generate_plan` function during the goal setting dialogue state. The preceding dialogue state prompts include explicit instructions to solicit relevant information and preferences from the user (e.g., goals, resources, injuries, past experience, preferred activities, and schedule), ensuring that the goal-setting process is collaborative, not prescriptive, and that the conversation history contains sufficient context for the `generate_plan` function to generate a personalized plan. We do not allow the agent to advance out of the goal setting dialogue state until the `generate_plan` function has been called. In absence of this manual check, we encountered hallucination issues with the agent telling the user it would generate a plan without calling the underlying function.

Moreover, we expose `add_workout` and `delete_workout` functions to the at-will chat agent, allowing the at-will agent to make direct edits to the user's current plan via conversational interaction. We do not provide the at-will agent access to the `generate_plan` function—early tests indicated that hallucination and tool use errors could lead to highly destructive edits or deletions. Making small, granular edits to the current plan proved to be more reliable than generating a replacement plan directly.

B Ambient Display Logic

Our ambient display encodings and logic were strongly inspired by UbiFit [28, 32] and modified to match our four-week study progression. The ambient display advances (i.e., a flower in the garden grows) as the user completes workouts in their weekly plan in 20% increments. For example, a flower grows if completing a workout advances the plan progress from 30% to 40% completion, but not from 20% to 30%. The flower fully blooms when the weekly plan is 100% completed. If the user completes their plan in week t, the flower remains in the garden and a new flower starts growing in week $t + 1$. If they do not complete their plan in week t, the flower starts growing again from scratch in week $t + 1$. In addition to a fully-bloomed flower, we add additional elements (week 2: bird on a branch; week 3: beehive on a branch; week 4: bird and birdhouse) to the display.

For each completed workout, a critter is drawn to the ambient display above the growing flowers. Users receive a bee for walking and a butterfly for other activites, with different colored buttlerflies for different activity types (red: cardio, orange: strength, green: team sports, yellow: flexibility & dance, blue: outdoor recreation, purple: misc). A small (< 15min), medium (15-30 min), or large (30+ min) critter is drawn depending on the exercise duration. To avoid cluttering the display, critters are reset each week. After the final screen (week 4, 100%), the ambient display remains fixed in subsequent weeks, but critters continue to appear.

C Field Study

In this section, we provide full details on our statistical analysis, custom survey questions, and complete survey results.

C.1 Statistical Analysis

Let $Y_{i,w,t}$ represent the observed outcome (step count, energy burned, exercise time, or distance walking/running) for participant i, measured on day $t \in \{1, \ldots, 7\}$ within study week $w \in \{0, \ldots, 4\}$ (week 0 aggregates all baseline weeks). Define an indicator variable for the study period, $S_w = \mathbf{1}\{w > 0\}$, and let T_i denote treatment assignment, where $T_i = 1$ for participants in the LLM condition and $T_i = 0$ for control. Let $\boldsymbol{D}_t$ be an indicator for the weekday (Monday–Saturday, with Sunday as reference). We specify the following three-level linear mixed-effects model, with days (t) nested within weeks (w) nested within participants (i):

$$
\begin{aligned}
Y_{i,w,t} = \; & \underbrace{\beta_0}_{\text{baseline mean}} + \underbrace{\beta_1 \cdot S_w}_{\text{H1: baseline–study period difference}} + \underbrace{\beta_2 \cdot S_w \cdot T_i}_{\text{H2: treatment–control difference in change}} \\
& + \underbrace{\beta_3 \cdot S_w \cdot w}_{\text{overall weekly trend}} + \underbrace{\beta_4 \cdot S_w \cdot T_i \cdot w}_{\text{H3: treatment–control difference in weekly trend}} \\
& + \underbrace{\boldsymbol{\gamma}^\top \boldsymbol{D}_t}_{\text{day-of-week effects}} + \underbrace{b_{0,i} + b_{0,i,w}}_{\substack{\text{random intercepts for} \\ \text{participant and participant–week}}} + \varepsilon_{i,w,t}
\end{aligned}
$$

Random intercepts $b_{0,i} \sim \mathcal{N}(0, \tau^2_{\text{person}})$ capture participant-level differences in mean activity. Random intercepts $b_{0,i,w} \sim \mathcal{N}(0, \tau^2_{\text{week}})$ capture week-to-week deviations within each participant. We additionally allow for heteroskedasticity across participants by modeling the residuals as $\varepsilon_{i,w,t} \sim \mathcal{N}(0, \sigma_i^2)$, where σ_i^2 is a participant-specific variance.

In this model, hypothesis tests correspond directly to fixed effects: coefficient β_1 tests whether average physical activity differs significantly between the baseline and study period (H1); coefficient β_2 assesses if the magnitude of this difference varies significantly between the treatment and control groups (H2); and coefficient β_4 evaluates whether the weekly rate of change (persistence) in activity over the study period differs significantly between groups (H3).

To account for systematic data missingness (e.g., due to privacy permissions, data upload errors, or participants forgetting to wear their watch), we excluded participants with fewer than 30 baseline days or with three or more missing study days from our analysis. This yielding a sample of 54 participants for both step count and distance walking/running (6,348 observations), 41 participants for energy burned (4,615 observations), and 36 participants for exercise time analyses (3,811 observations). Notably, relaxing our exclusion criteria and including all participants in our model does not change the directionality or significance of our effects.

We explored a variety of models varying in complexity, from two-sample t-tests and standard repeated-measures ANOVA, to more sophisticated mixed-effects models including multiple levels, random slopes, and non-linear time trends. We also examined alternative coding schemes (e.g., treating weeks or days as factors) and aggregating data at the weekly rather than daily level. We selected this final model because it provided the best statistical fit to the data (lowest AIC/BIC), leveraged daily-level observations to maximize statistical power, and avoided the strong assumption that treatment effects vary linearly at the daily level. Instead, the model captures treatment effects as linear weekly trends, aligning well with our hypothesis concerning persistence (H3). Additionally, this approach explicitly accounts for heteroskedasticity by allowing participant-specific residual variances, acknowledging that observed changes in activity are more noteworthy for participants whose baseline behavior was less variable. Importantly, the patterns of significance, directionality of effects, and substantive conclusions drawn from our results remained consistent across all modeling approaches tested.

C.2 Survey Measures

Physical Activity & Health: General Assessment
In general, would you say your health is …
How important is health to you personally?
In general, how would you rate your physical fitness?
During the Physical Activity st 7 days, how much exercise did you get?
Physical Activity & Health: Satisfaction
How satisfied are you with your current level of physical activities?
How satisfied are you with your current physical health?
How satisfied are you with your current mental health?
How satisfied are you with your current life situation in general?
Physical Activity & Health: Motivation
How much do you intend to increase your current level of physical activity?
How motivated are you to adopt a healthier lifestyle?
User Experience & Advice Quality (modified from [55])
I received personalized physical activity advice.
I received actionable physical activity advice.
I received generic physical activity advice.
I felt comfortable sharing my concerns.
I felt supported in my physical activity.
I feel capable of overcoming challenges.
I feel more motivated to change.
I was asked for my opinion about what activities I would like to do.
I felt like my unique situation and concerns were understood.
I received unsolicited advice.
The system was empathetic.
The system used my data in a way that was relevant.
The system helped me identify obstacles to engaging in physical activity.
The system helped me reflect on what motivates me to be physically active.
The system helped make my own ideas about how to increase my physical activity more specific.
Interacting with the system provided me with new insights about my physical activity.
Anthropomorphism (Post-only)
Interacting with the app felt like interacting with a human.
Interacting with the app made me feel like my wellbeing was cared for.
Beebo cared about my wellbeing.

Table 8: Full wording of every custom question included in the onboarding/offboarding survey.

Measure	Treatment (LLM) Pre	Post	Δ_T	Control (no LLM) Pre	Post	Δ_C	$\Delta_T - \Delta_C$
Stage of Change [78] *(1: pre-contemplation, 5: maintenance)*	1.88 (0.59)	2.85 (0.73)	0.96 (0.96)	1.89 (0.74)	2.61 (1.07)	0.71 (1.01)	0.25
IPAQ [33] *(MET-min)*	415.31 (329.00)	749.73 (467.88)	334.42 (505.20)	548.18 (637.10)	893.96 (741.87)	345.79 (965.50)	-11.36
IPAQ [33] *(1: Low, 3: High)*	1.27 (0.45)	1.77 (0.51)	0.50 (0.65)	1.25 (0.44)	1.68 (0.67)	0.43 (0.79)	0.07
Self-Efficacy [112]	2.83 (0.67)	2.82 (0.88)	-0.01 (0.93)	2.61 (0.76)	2.63 (0.86)	0.02 (0.99)	-0.03
Adequacy Mindset [141]	2.39 (0.94)	3.59 (0.74)	1.21 (0.78)	2.59 (0.85)	3.35 (0.93)	0.76 (0.85)	0.44
Process Mindset [16]	2.71 (0.61)	3.04 (0.77)	0.33 (0.48)	2.62 (0.82)	2.84 (0.74)	0.22 (0.48)	0.11
Barriers to Being Active [110]	2.59 (0.56)	2.35 (0.63)	-0.24 (0.36)	2.64 (0.62)	2.38 (0.71)	-0.27 (0.64)	0.02
SUS [19] *(0–100)*	93.85 (8.89)	82.21 (20.85)	-11.63 (18.64)	93.75 (8.29)	89.02 (8.91)	-4.73 (8.80)	-6.90
SASSI [50]	4.40 (0.43)	4.07 (0.75)	-0.33 (0.63)	4.38 (0.36)	4.25 (0.44)	-0.13 (0.46)	-0.20
TSRI [11]	3.80 (0.58)	3.85 (0.86)	0.06 (0.77)	3.93 (0.53)	3.61 (0.92)	-0.33 (0.80)	0.38
eHealth Literacy [96]	4.46 (0.50)	4.35 (0.82)	-0.11 (0.80)	4.44 (0.53)	4.59 (0.52)	0.15 (0.39)	-0.26
(Adapted) SDM-Q-9 [67]	4.06 (0.64)	4.18 (0.79)	0.12 (0.46)	3.75 (0.85)	3.48 (0.92)	-0.26 (0.69)	0.38
Physical Activity & Health – General Assessment (Custom)							
Overall Average	2.93 (0.47)	3.31 (0.53)	0.38 (0.39)	2.69 (0.53)	3.01 (0.60)	0.32 (0.46)	0.06
General Health Assessment	2.81 (0.69)	3.15 (0.61)	0.35 (0.49)	2.75 (0.80)	2.93 (0.81)	0.18 (0.55)	0.17
Health Importance	4.04 (0.60)	4.15 (0.61)	0.12 (0.52)	3.71 (0.76)	3.86 (0.85)	0.14 (0.71)	-0.03
Physical Fitness Rating	2.12 (0.77)	2.38 (0.80)	0.27 (0.67)	1.96 (0.84)	2.25 (0.84)	0.29 (0.71)	-0.02
Exercise Past 7 Days	2.69 (0.71)	3.62 (0.70)	0.92 (0.91)	2.19 (0.66)	3.00 (0.85)	0.81 (0.91)	0.11
Physical Activity & Health – Satisfaction (Custom)							
Overall Average	2.15 (0.42)	2.64 (0.49)	0.49 (0.39)	2.19 (0.50)	2.53 (0.57)	0.33 (0.40)	0.15
Satisfaction: PA Level	1.82 (0.81)	3.23 (1.01)	1.41 (0.93)	1.95 (0.71)	2.98 (1.00)	1.02 (0.97)	0.39
Satisfaction: Physical Health	2.28 (0.96)	3.15 (1.04)	0.87 (1.06)	2.52 (1.10)	3.17 (1.05)	0.64 (0.93)	0.23
Satisfaction: Mental Health	3.38 (0.93)	3.82 (0.99)	0.44 (0.80)	3.36 (1.11)	3.50 (1.21)	0.14 (0.91)	0.29
Satisfaction: Life Situation	3.44 (0.94)	3.64 (0.95)	0.21 (0.77)	3.33 (1.10)	3.52 (1.03)	0.19 (0.44)	0.01
Physical Activity & Health – Motivation (Custom)							
Overall Average	3.50 (0.57)	3.48 (0.70)	-0.02 (0.61)	3.18 (0.78)	3.38 (0.82)	0.20 (0.72)	-0.22
Intention to Increase PA	3.31 (0.62)	3.23 (0.71)	-0.08 (0.69)	3.04 (0.79)	3.07 (0.86)	0.04 (1.04)	-0.11
Motivation Healthier Lifestyle	3.69 (0.68)	3.73 (0.83)	0.04 (0.72)	3.32 (1.02)	3.68 (1.09)	0.36 (0.73)	-0.32
User Experience & Advice Quality (drawn from [55])							
Overall Average	4.40 (0.47)	4.04 (0.79)	-0.37 (0.52)	4.12 (0.54)	3.63 (0.74)	-0.50 (0.74)	0.13
Actionable Advice	4.81 (0.40)	4.42 (1.03)	-0.38 (0.80)	4.71 (0.53)	4.07 (0.98)	-0.64 (1.03)	0.26
Personalized Advice	4.58 (0.58)	4.23 (0.99)	-0.35 (1.02)	3.82 (0.98)	4.00 (1.25)	0.18 (1.33)	-0.52
Generic Advice	2.19 (1.17)	2.58 (1.17)	0.38 (1.50)	2.79 (0.99)	2.79 (1.17)	0.00 (1.22)	0.38
Comfortable Sharing Concerns	4.73 (0.53)	4.23 (0.99)	-0.50 (0.76)	4.75 (0.52)	3.96 (0.92)	-0.79 (0.99)	0.29
Feel Supported	4.46 (0.76)	4.23 (0.99)	-0.23 (0.91)	4.39 (0.69)	4.14 (0.97)	-0.25 (0.93)	0.02
Feel Capable of Overcoming Challenges	4.35 (0.69)	3.85 (1.19)	-0.50 (0.95)	4.11 (0.88)	3.54 (1.17)	-0.57 (1.20)	0.07
Feel Motivated to Change	4.54 (0.58)	3.96 (1.37)	-0.58 (1.14)	4.46 (0.64)	3.89 (1.10)	-0.57 (1.32)	-0.01
Asked for my Opinion	5.00 (0.00)	4.81 (0.40)	-0.19 (0.40)	4.46 (0.84)	3.46 (1.57)	-1.00 (1.78)	0.81
Understood Unique Situation & Concerns	4.38 (0.75)	3.88 (1.24)	-0.50 (1.30)	3.43 (1.14)	2.71 (0.98)	-0.71 (1.54)	0.21
Unsolicited Advice	1.42 (0.81)	1.96 (1.11)	0.54 (1.33)	2.18 (1.33)	1.86 (1.08)	-0.32 (1.61)	0.86
Empathetic	4.23 (0.99)	4.12 (0.99)	-0.12 (0.82)	3.54 (1.04)	2.79 (1.03)	-0.75 (1.24)	0.63
Used Data in a Relevant Way	4.65 (0.63)	4.27 (0.96)	-0.38 (0.98)	4.50 (0.88)	4.36 (0.78)	-0.14 (1.11)	-0.24
Helped Identify Obstacles	4.23 (0.91)	3.73 (1.12)	-0.50 (1.07)	4.36 (0.95)	3.11 (1.31)	-1.25 (1.43)	0.75
Helped Reflect on Motivation	4.08 (1.09)	3.65 (1.20)	-0.42 (1.39)	4.57 (0.74)	3.43 (1.32)	-1.14 (1.18)	0.72
Helped Make Own Ideas More Specific	4.35 (1.06)	3.92 (1.13)	-0.42 (1.03)	4.07 (0.94)	3.68 (1.19)	-0.39 (1.17)	-0.03
Provided New Insights	3.65 (1.16)	3.81 (1.13)	0.15 (1.19)	3.79 (1.17)	3.54 (1.35)	-0.25 (1.29)	0.40
Anthropomorphism (Post-only, Custom)							
Overall Average	–	3.51 (1.17)	–	–	2.83 (0.98)	–	–
Human-Like Interaction	–	2.88 (1.37)	–	–	1.89 (1.20)	–	–
Wellbeing Cared For	–	3.85 (1.22)	–	–	3.25 (1.21)	–	–
Beebo Cared	–	3.81 (1.41)	–	–	3.36 (1.06)	–	–

Table 9: Pre/Post Survey Results. We report mean (SD) survey measures for participants in the treatment (LLM) and control (no LLM) conditions for pre- and post-study surveys. Δ refers to the mean (SD) of participants' post-pre difference in scores within each condition while $\Delta_T - \Delta_C$ refers to the treatment-control difference in post-pre differences. For all measures except stage of change, IPAQ, and SUS, responses are standardized to fall on a 5pt-Likert scale (1: Strongly disagree; 5: Strongly agree). For all surveys except Barriers to Being Active (where lower scores indicate lower perceived barriers to activity), responses are coded such that higher scores are more desirable.

Question	Treatment (LLM)					Control (no LLM)				
	Mean	Day 1	Day 28	Δ_T	β_T	Mean	Day 1	Day 28	Δ_C	β_C
I am satisfied with my current level of physical activity.	3.17 (1.09)	2.59 (1.18)	3.45 (1.01)	1.17 (0.99)	0.022	3.06 (1.07)	2.71 (1.10)	2.91 (1.16)	0.43 (1.22)	0.006
I am a physically active person.	3.15 (1.03)	2.77 (1.02)	3.32 (0.95)	0.61 (0.78)	0.030	3.02 (1.05)	2.53 (0.87)	3.00 (1.00)	0.71 (0.99)	0.009
I am committed to my physical activity goal.	4.27 (0.66)	4.23 (0.53)	4.09 (0.81)	0.17 (0.38)	-0.007	3.96 (0.84)	3.82 (0.73)	3.91 (0.90)	-0.21 (1.31)	0.000
How hopeful are you about your physical health today?	4.07 (0.70)	3.91 (0.68)	4.14 (0.71)	0.44 (0.78)	0.005	3.84 (0.85)	3.88 (0.60)	3.87 (0.87)	-0.21 (0.80)	-0.003
How was your mood today? (unpleasant/pleasant)	3.93 (0.80)	4.14 (0.64)	4.32 (0.84)	0.33 (0.91)	0.004	3.91 (0.83)	4.00 (0.71)	3.96 (0.77)	-0.21 (0.97)	0.002

Table 10: Daily Survey Results. We report mean (SD) values by condition across all days, for day 1, and for day 28. Δ is the change from day 1 to day 28 (mean difference, SD). β is the per-day slope from a linear regression of score on day. All responses map to a 5 point Likert scale.

Question	Treatment (LLM)					Control (no LLM)				
	Mean	Week 1	Week 4	Δ_T	β_T	Mean	Week 1	Week 4	Δ_C	β_C
Overall App Rating	3.84 (0.74)	3.90 (0.67)	3.98 (0.81)	-0.05 (0.63)	-0.01	4.02 (0.69)	3.94 (0.68)	4.02 (0.68)	0.30 (0.69)	0.04
Visual Wallpaper Rating	3.78 (1.12)	3.70 (1.18)	3.93 (1.22)	0.29 (0.61)	0.08	3.94 (0.94)	3.73 (0.98)	4.14 (0.81)	0.24 (0.82)	0.15
Exercising is… (difficult/easy)	3.11 (0.82)	2.92 (0.78)	3.35 (1.11)	0.39 (1.03)	0.14	2.93 (1.00)	2.90 (1.17)	2.87 (0.98)	0.00 (0.50)	0.02
Exercising is… (boring/fun)	3.17 (0.91)	3.13 (0.86)	3.27 (1.03)	0.24 (0.70)	0.04	2.96 (1.06)	2.81 (1.10)	2.87 (1.10)	0.09 (0.94)	0.06
Barriers To Being Active (4 item subset)	2.54 (0.70)	2.51 (0.74)	2.53 (0.71)	-0.02 (0.66)	-0.01	2.79 (0.85)	2.92 (0.86)	2.60 (0.96)	-0.44 (0.39)	-0.06

Table 11: Weekly Survey Results. We report mean (SD) values by condition across all weeks, for week 1, and for week 4. Δ is the change from week 1 to week 4 (mean difference, SD). β is the per-week slope from a linear regression of score on week. All responses were rescaled to map to a 5 point Likert scale.