PlanFitting: Personalized Exercise Planning with Large Language Model-driven Conversational Agent

Donghoon Shin*
dhoon@uw.edu
University of Washington
Seattle, WA, USA

Gary Hsieh
garyhs@uw.edu
University of Washington
Seattle, WA, USA

Young-Ho Kim
yghokim@younghokim.net
NAVER AI Lab
Seongnam, Gyeonggi, Korea

ABSTRACT

Creating personalized and actionable exercise plans often requires iteration with experts, which can be costly and inaccessible to many individuals. This work explores the capabilities of Large Language Models (LLMs) in addressing these challenges. We present PlanFitting, an LLM-driven conversational agent that assists users in creating and refining personalized weekly exercise plans. By engaging users in free-form conversations, PlanFitting helps elicit users' goals, availabilities, and potential obstacles, and enables individuals to generate personalized exercise plans aligned with established exercise guidelines. Our study—involving a user study, intrinsic evaluation, and expert evaluation—demonstrated PlanFitting's ability to guide users to create tailored, actionable, and evidence-based plans. We discuss future design opportunities for LLM-driven conversational agents to create plans that better comply with exercise principles and accommodate personal constraints.

ACM Reference Format:
Donghoon Shin, Gary Hsieh, and Young-Ho Kim. 2025. PlanFitting: Personalized Exercise Planning with Large Language Model-driven Conversational Agent. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 8–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3719160.3736607

1 INTRODUCTION

Despite the benefits of regular exercise, people often struggle to meet recommended physical activity guidelines [21, 70, 72]. To facilitate regular exercise, many digital tools follow the approaches of personal informatics [50], including activity tracking [23, 29, 36], visualization [7, 17, 18, 42, 50, 51], and self-reflection [14, 42, 46] on activity data. However, existing tools place less focus on *exercise planning*, leaving users to create their own workout schedules, which can be challenging without domain expertise, especially when tailoring it to personal lifestyle constraints [19, 75]. As a result, people rely on professional planners (*e.g.*, personal trainers, medical practitioners); yet involving experts also presents various setbacks, such as high costs, inaccessibility, and lack of customization due to broad client bases [11, 57, 59, 64]. To enhance personalization in exercise planning, a handful of works have explored crowdsourced and peer-supported planning [5, 6], but these approaches still require human effort, time, and financial investment, while also depending on users to articulate clear preferences upfront—often challenging without expert input.

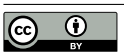
One potential solution to tackle these challenges is to use LLM-driven conversational agents (CAs) to tailor exercise plans to individuals. With the availability and scalability that CAs offer, we posit that CAs driven by LLMs can guide users to create and continuously refine plans tailored to their individual contexts. Specifically, recent advances suggest the potential of LLM-driven CAs in collecting information for social needs (*e.g.*, healthcare [45]) and synthesizing information in the knowledge task (*e.g.*, [12, 47, 55]), through iterative turn-taking with the user. Highlighting these potentials, in this work, we explore how LLM-driven CAs can be used to help individuals craft and revise personalized exercise plans.

To that end, we first conducted formative interviews exploring current practices in creating personalized exercise plans and the challenges faced in professional planning contexts. From the interviews with professional exercise planners ($N = 5$) and lay individuals (*i.e.* clients; $N = 8$) who have experience in setting up personalized exercise plans with planners, we characterized key steps in crafting personalized exercise plans—goal-setting, collecting availabilities and anticipated obstacles, prescribing plans, and iteration, while grounding the plans on the core high-level guidance suggested by existing exercise guidelines (*e.g.*, ACSM [3, 28]). Additionally, we found that planners often face difficulties integrating exercise prescriptions into the irregular schedules of clients, with limited incorporation of client input during the iterative process of revising plans.

Based on these insights, we designed and developed PlanFitting, an LLM-driven CA that assists lay individuals in creating and refining their personalized exercise plans grounded on guidelines through interactive dialogue. With a dynamic prompting approach, PlanFitting leads users to engage them in conversations that gather essential information about their constraints identified in our formative study (*i.e.*, exercise goals, availabilities, and potential obstacles to adherence). Using this information, the agent recommends exercises from the dataset through the retrieval-augmented generation and presents the plan in the form of implementation intention (*i.e.*, `IF-THEN` rules) [30]—a concise, flexible scheduling framework grounded in behavioral psychology that links user intentions to specific events without rigid time scheduling [15], while aligning the plan with established exercise guidelines [3].

We conducted a user study ($N = 18$) where the participants formulated a weekly plan and refined it with PlanFitting. Our results found that PlanFitting effectively helped participants articulate personalized constraints while adapting to their unique chatting styles. Also, participants found PlanFitting to be useful and usable,

*Donghoon Shin conducted this work as a research intern at NAVER AI Lab.



CUI '25, July 8–10, 2025, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/2025/07.
https://doi.org/10.1145/3719160.3736607

and highlighted the agent's role in guiding them towards creating personalized and actionable plans. Additionally, our intrinsic evaluation revealed that the generated plans reliably followed the established exercise guidelines, and expert planners ($N = 3$) who evaluated the generated plans based on the exercise principle (*i.e.*, FITT [20]) evaluated the *frequency*, *intensity*, and *time* composition of the generated plans to be above average. However, they also identified opportunities to enhance the combination of exercise *types*. Based on qualitative feedback from participants and expert planners, we also explore design implications for improving the use of LLM-driven CAs in creating personalized exercise plans.

The main contributions of our work, along with the corresponding sections in the paper, are as follows:

- We present the results of our formative study, revealing the process and challenges of exercise planning between clients and expert planners, which informed the design of our conversational agent (§3);
- We introduce PlanFitting, an LLM-driven conversational agent that assists users in creating and refining personalized exercise plans. We present the agent's operationalization—including dialogue management and the interaction between the conversational agent and user—towards creating personalized and guideline-informed exercise plans (§4);
- We present empirical findings from (i) a user study exploring how users interact with and perceive PlanFitting, (ii) an intrinsic evaluation assessing how well the generated plans follow established exercise guidelines, and (iii) an expert evaluation assessing the quality of the generated plans (§5, §6).

2 RELATED WORK

Our work builds on prior research that explored (1) personalized exercise planning, (2) technology-mediated exercise support and planning, and (3) LLM-driven conversational agents.

2.1 Crafting Personalized and Actionable Exercise Plans

Engaging in regular physical activity is essential for a healthy lifestyle; however, many people struggle to integrate sufficient exercise into their daily lives [19, 75]. To address this, establishing and adhering to exercise plans has proven effective for motivating individuals to sustain consistent physical activity [30, 54]. In response, several evidence-based guidelines have been proposed; for instance, the American College of Sports Medicine (ACSM) [3, 28] and the U.S. Department of Health and Human Services [58, 65] have created widely accepted guidelines that health professionals utilize to formulate effective exercise regimens. These guidelines provide general recommendations for planning (*e.g.*, advising a minimum of 150 minutes of moderate-intensity exercise per week) [3, 28, 58, 65], along with definitions of exercise-related terms. With these comprehensive guidelines, individuals can further tailor their exercise plans to their preferences and constraints, which is known to contribute to the successful adoption of plans, though achieving such personalization is not trivial and often requires the expertise of exercise professionals.

Another line of research in behavioral psychology and sports medicine has explored the effective intervention and format of exercise prescriptions. One well-known approach is *implementation intention*, which comprises a specific plan linking a particular circumstance to corresponding actions [30, 34]. Formatted as `IF-THEN` rules, implementation intentions are often combined with action planning by including environmental cues [34]. For example, one can set up an exercise plan like "`IF` *I come back home in the evening*, `THEN` *I will jog for 30 minutes.*" By effectively transforming intentions into actionable steps, implementation intentions have demonstrated success in various behavior change contexts (*e.g.*, managing a healthy diet [2, 4, 33, 66], reducing bedtime procrastination [71], smoking cessation [16, 56]). Likewise, in the context of exercise, implementation intentions have been shown to be effective in promoting physical activities [52], suggesting its adaptability to personalized exercise planning.

2.2 Technology-mediated Exercise Support and Planning

Given the importance and barriers of regular exercise, researchers in HCI have long investigated the design of digital tools to facilitate the tracking of physical activity [7, 17, 18, 40–43, 46, 51]. These tools commonly incorporated a personal informatics and self-tracking approach, where the tool provides insights about the user's progress and status of exercise so that they stay motivated and knowledgeable about themselves [50]. For example, UbiFit Garden employed metaphoric visualization of various daily exercise metrics (*i.e.*, exercise categories and amount) to help users keep up with the progress of activity at a glance [18]. Reflection Companion engages users in a daily SMS dialogue that promotes self-reflection on their physical activity levels captured by activity trackers [46].

Meanwhile, research on technology support for exercise planning is relatively sparse, with only a few works exploring planning tools (*e.g.*, [5, 6, 48, 75]). Xu *et al.* investigated digital planning experiences for physical activities [75], and Lee *et al.* probed the effect of reflective strategies in physical activity planning [48]. Agapie *et al.* proposed involving peers and crowdworkers to help generate custom exercise plans for health behavior change [5, 6]. These works showed technology's potential in constructing personalized exercise plans, yet they still require substantial human involvement for plan formulation which limits the sustainability and scalability. Although a handful of commercial applications (*e.g.*, [24, 25]) have attempted to use AI for planning resistance training, they are limited to maximizing the effectiveness of muscle growth, and lack adaptability to individual schedules. To address these limitations, our work proposes leveraging the scalability of conversational agents by exploring their use in supporting iterative planning to assist users in creating personalized exercise plans.

2.3 Leveraging LLM-driven Conversation Agents for Personalized Exercise Planning

Conversational agents (CAs) have found widespread use in gathering information for various social purposes, such as web surveys [39] and fostering self-disclosure [49, 63], as they are easily scalable and readily available to use. Their applications extend to healthcare and well-being contexts, where prior works

in healthcare leveraged CAs to collect health-related information from users [9, 22], and have been shown to be a preferred way of providing social needs information—especially for those with lower health literacy [44, 45]. However, conventional rule-based CAs are typically known to suffer from constrained interaction capabilities, lack of extensibility to other domains once designed, and rigid input demands, particularly because they lack robust natural language adaptability and comprehension [1, 37, 74]. Moreover, analyzing the information gathered by these agents to compose meaningful insights still relies heavily on human practitioners, resulting in substantial manual efforts.

Recent advancements in LLMs demonstrate the potential of CAs to engage more fluidly in dialogue while synthesizing information to deliver valuable insights in real-time. By incorporating detailed behavioral guidelines (*i.e.*, preprompt/system prompt), these agents can adapt their conversational style to suit diverse contextual requirements, enabling open-ended interactions without the need for extensive training dialogue corpora. This mechanism has streamlined bootstrapping in novel conversational topics, as evidenced by the broad range of applications from general-purpose agents (*e.g.*, ChatGPT [60], Gemini [32]) to specialized research prototypes (*e.g.*, health data collection [74], recommender system [13, 26]). Moreover, innovations in LLM frameworks (*e.g.*, LangChain [47], AgentVerse [12]) and applications (*e.g.*, [55]) have further demonstrated how LLM-driven agents can be harnessed to perform complex, knowledge synthesis tasks (*e.g.*, data-driven question-answering). In the context of exercise planning, this suggests the possibility of harnessing LLM-driven CAs to not only flexibly collect individual users' constraints but also to integrate and analyze these to create cohesive plans.

Despite these, ensuring an LLM has learned specific knowledge during its pretraining remains challenging [38]. This has been pointed out to make LLM-based CAs prone to returning errors, particularly regarding domain-specific conversations that require specialized knowledge [27, 38, 68]–including exercise planning. For example, when tasked with generating exercise plans, the CA may offer recommendations for exercise types and amounts that lack evidence, especially with regard to an individual's unique situation. One potential approach to enhance the accuracy and credibility of such conversations is Retrieval-Augmented Generation (RAG), in which critical knowledge required for the task is retrieved from an external knowledge base and incorporated into the preprompt to augment the LLM agent's response generation [27, 67].

Drawing inspiration from research on using CAs for social needs and the adaptability of LLMs, this work investigates how LLM-driven CAs can be designed to interact with users towards the creation of personalized plans. More specifically, we aim to enable the free-form expression of user constraints and requirements for exercise planning, and synthesize them into evidence-based plans. To enhance the robustness of following up the dialogue context, we also propose a design choice to implement a separate LLM routine for the agent that generates a dialogue summary, which is injected into the preprompt of the LLM for conversation. Lastly, we incorporate the retrieval mechanism by making the agent refer to an external exercise database to avoid recommending seemingly plausible but irrelevant (*i.e.*, "hallucinated" [27]) exercise types. In our work, we demonstrate how these mechanisms can holistically support personalized exercise planning that is reliable and grounded on credible knowledge.

3 FORMATIVE STUDY

To understand the current practice of conducting personalized exercise planning and the challenges that arise during the process, we conducted a formative interview study with exercise planners ($N = 5$) and clients ($N = 8$). The study protocol was reviewed and approved by the institutional review board.

Exercise planners. From an in-house clinic and a corporate internal network, we recruited five experts (FP1–5; three females and two males) who are experienced in setting up personalized exercise plans for clients. Of all, three were physical therapists, another was a physiatrist, and the other was a kinesiologist. On average, they had 9.8 years ($SD = 4.5$) of experience in advising and planning exercise planning.

Clients. We recruited eight individuals (FC1–8; 6 females and 2 males) by advertising our study on a local community platform and the corporation's internal bulletin boards. We required participants to have experience setting up their personalized exercise plans under the advice of exercise experts (*e.g.*, clinicians, physical therapists, personal trainers, etc.). Clients were aged between 26 and 45 ($M = 35.0$); three participants responded that they have/had engaged in exercise under the personalized exercise plans for less than three months, three participants for 3 to 6 months, and the other two participants for more than six months.

We invited each participant to a 1-hour semi-structured interview session. During the session, we asked each exercise planner to primarily share insights into (1) their planning procedures for clients and (2) the challenges they encountered while setting up personalized plans for/with the clients. Likewise, clients were prompted to elaborate on (1) their experiences and process of planning exercises with exercise planners and (2) the challenges they faced during the planning. Each interview was audio-recorded and later transcribed, and we compensated 50,000 KRW (approximately 35 USD) and 30,000 KRW (approximately 21 USD) for each planner and client, respectively.

After the interview, we analyzed the interview transcripts using thematic analysis [10]. The analysis was done in a bottom-up approach, where the two authors first familiarized themselves with the raw responses independently. Then, each author identified emerging themes from the responses, brought these themes to a regular meeting, and compared the themes until they reached a consensus. As a result, we could derive the final themes as detailed in Section 3.1 and 3.2.

3.1 Practice of Personalized Exercise Planning

First, planners reported that they primarily inform the exercise plans with globally recognized guidelines (*e.g.*, ACSM guidebook [3]), which emphasizes engaging in a minimum of 150 minutes of moderate-intensity exercise per week. However, they suggested that these guidelines do not provide specific guidance on tailoring to individuals' varying lifestyles: *"Actually, even if you take a look at those exercise planning guidebooks, there won't be anything more detailed than [showing a page that defined some case studies of individuals]*

(...) that's the end of 'evidence-based' personalization." (FP4) As such, planners use them as a flexible framework rather than strict rules, making tailored modifications while adhering to such high-level principles: *"I'm just following a broad guide and customizing a lot in that scope. Shouldn't the details within it be personalized?"* (FP2)

More specifically, we could characterize the process of personalization, and surface the common information that planners gather from clients in this process to tailor plans to their lifestyles, such as personal goals for the exercise, personal obstacles, and feedback (during the follow-up sessions), delivered through either verbal communication or a combination of a survey form and oral report:

Understanding client's main goals for exercise. Every planner responded that they begin by identifying the client's goal for the exercise, highlighting the importance of defining the purpose and setting clear objectives to motivate clients. To enable this, they engaged in conversations with clients to find out their own necessity and benefits of exercise to enhance motivation, particularly for newcomers: *"For managing exercise plans, it's crucial to first motivate by discussing goals first rather than just telling them to do it."* (FP1)

Surfacing available amount of times for exercise and potential obstacles. Once identifying the goals for the exercise, planners are reported to ask clients questions about their availabilities, such as how much time they would be available to spare for exercise: *"For those who don't have set regular office hours or for nurses working 3/4 shifts, I ask and look at how much personal time the client can exercise on a regular basis."* (FP3) Also, planners surface factors from clients that may potentially make it challenging for them to exercise during those times (*e.g.*, physical constraints, parenting), to make the exercise planning more viable and realistic: *"I told my planner when my menstrual cycle comes (...) And (as a developer in a company) I told them whenever there is a schedule for releasing a new version that my condition won't be good for about three following days."* (FC5)

Prescribing plans. Based on collected exercise goals, availabilities, and obstacles, planners create a personalized exercise plan for clients. While planners are willing to provide detailed plans down to specific times, the limited availability of planners makes this approach impractical: *"I can't do detailed time planning (...) It seems inconsistent (with my current availability) to generate highly detailed plans, like scheduling at a certain time."* (FP4) As a result, planners and clients typically receive a weekly exercise plan with recommended days and hours, exercise types, allowing clients to exercise at their own convenience to meet their requirements: *"They (planner) didn't ask me to exercise at a specific time; they just told me to do a certain amount of some exercises during the week."* (FC4)

Revisit regularly (e.g., weekly, bi-weekly) to share feedback and iterate on the plan. Emphasizing the importance of viewing the exercise planning as a feedback-driven iteration, rather than a one-time interaction, planners and clients revisit the plans regularly (*e.g.*, weekly, bi-weekly) to check if the exercises need to be modified: *"There are types of exercise that go in and out (...) After solving the urgent problem, if I wanna get a nicer body shape, other exercises may go in or out."* (FC1) Gathering newly emerged feedback and constraints, planners make adjustments to exercise types and/or duration: *"Clients first give it a try, and I gather feedback when they come back in the following week based on their experience trying the exercise plan. If they think it won't work for any reason, I ask them to let me know, and we can start the revisions from there, just like forming and iterating on a hypothesis."* (FP2)

3.2 Challenges of Personalized Exercise Planning

3.2.1 Difficulty of contextualizing the exercise within their own schedule. After the prescription of a broadly defined weekly exercise plan, clients are required to incorporate these exercises into their own schedules by themselves. However, clients from our interviews reported that such an 'autonomous' process, without clearer support on identifying when to exercise within their actual schedule, makes it difficult for them to cope with unexpected variables (*e.g.*, appointments, work schedules). As such, adhering to the plans becomes highly reliant on their own motivation, making clients prone to becoming complacent: *"I think it's mostly about getting the number of exercises and then performing them on my own, so my own willingness is the most important factor (...) If I suddenly have to work at night, I just end up not doing exercise that day because there's no one pushing me to do and I feel like I can just do it later."* (FC2)

In particular, these issues are reported to worsen over time. As time passes, various triggers that may lower motivation are reported to emerge, such as moments of stagnation during their exercise progress, which is exacerbated over time, leading to a tendency to continuously postpone or skip prescribed exercise: *"If you aim for a weight loss, there are times when you reach a point where you're not losing any more weight (...) then my motivation decreased a bit, so sometimes I took a day or two off, rested a bit more, or skipped it in various other ways. So, I'm skipping more than I did in the beginning."* (FC6)

3.2.2 Limited availability of planners affecting the iteration process and adaptation to fluctuating schedules. Clients expressed struggles around accommodating sudden, unexpected time changes caused by their irregular lifestyles and work schedules. In these cases, reaching out to planners for real-time schedule adjustments is unrealistic, due to the other personal/work commitments planners have. As a result, clients reported their desire for more flexibility in communication when iterating on exercise plans: *"I have been meeting my planner every week (...) it was sad to see whenever I have a schedule change and need an alternative, I couldn't ask about the plan iterations right away for the other days."* (FC6)

Such an issue, ironically, is reported to make the whole exercise schedule of clients even more dependent on the planners' decision-making process. Consequently, if the weekly meeting is canceled as either the client or the planner is unable to attend the weekly meeting, it often results in a disruption of the exercise for the entire week: *"There were instances when the trainers were not available due to their other commitments (...) the whole exercise for the following week messed up."* (FC5)

3.2.3 Limited adaptability of planners in engaging with and incorporating client feedback. Even when meeting to discuss the exercise

plan, clients often struggle to have their concerns and input incorporated into the plans. Indeed, clients shared several anecdotes when they felt their opinions were dismissed, or they had to spend a considerable amount of time advocating for their points to be considered: *"You know, I can't see the planner every day and have to meet them face to face, and my daily conditions are different every day (...) but I always had to follow the same fixed program. I once went on a trip to [an attraction], but even when I explained this situation in advance my planner just asked me to keep exercising while traveling. It's too inflexible and feels too coercive."* (FC2)

The prescribed plan's inability to cater to unique constraints such as travel schedules could discourage clients from following through. In the worst case, disagreements stemming from this lack of flexibility have sometimes even led clients to discontinue their programs entirely: *"I and planners had disagreements on the types of exercise, and I discontinued planning for the exercise with my personal trainer from that moment."* (FC5)

4 PLANFITTING

Our formative study revealed the overall planning process of personalized exercise plans, as well as the challenges that emerge during the process. Building on these insights, we designed and implemented PlanFitting, a conversational agent system aimed to help individuals set up their personalized exercise plan and iterate on it. Focusing on the expressivity and comprehensibility that LLMs offer, we designed our system using LLMs to foster engaging interaction, while adapting to the unique constraints of users and allowing them to iterate their plans.

Informed by the planning procedure we surfaced from our preliminary study, we formulated the interaction process of PlanFitting into the following three stages: First, (1) the user provides exercise-related constraints (*i.e.*, goals, availabilities, obstacles) to the agent. Then, the agent (2) offers a personalized exercise recommendation based on the provided constraints and (3) generates a personalized weekly plan. Lastly, (4) the user may revisit PlanFitting where the agent assists in refining the plan by accommodating the user's changing constraints. This way, we aimed to accommodate PlanFitting to general user groups by taking into account their constraints around performing exercises, comparing them with the list of exercises and the strength/relevant muscles involved, and having LLM associate them and recommend feasible exercises. In the following, we describe the design of PlanFitting's dialogue system with the underlying LLM pipeline, as well as its implementation.

4.1 Interaction and Conversational UI Design

PlanFitting is designed as a web-based conversational UI (Figure 1), where the users can primarily interact with the CA on the chat panel (Figure 1-Chat panel) via natural language. It is supported by the dashboard providing an overview of the current status of the conversation (Figure 1-Dashboard), summarizing the exercise goal and constraints (Figure 1-Ⓐ), recommended exercise list from the system (Figure 1-Ⓑ), and the exercise plans (Figure 1-Ⓒ). Every information on the dashboard is automatically updated on every conversational turn so that the user can stay on track. In the following, we describe the detailed interaction flow between the CA and the user.

4.1.1 Collecting exercise-related user constraints. The CA first takes the lead by collecting essential information required for crafting a personalized exercise plan. Specifically, the CA proactively asks questions aimed at gathering the personal constraints of the user as follows:

(1) Exercise goals: The user's goal of exercise, either in a format of intended purpose or the specific muscle group they aim to target
(2) Availability: The user's available times for the exercise, either in the exact time format (*e.g.*, '*7 pm*') or in a descriptive form (*e.g.*, '*after work*')
(3) Potential obstacles: Any expected obstacles they anticipate that could potentially impede their exercise routine (*e.g.*, '*chance of working until late night*')

4.1.2 Exercise type recommendation. After the user has shared all the necessary constraints, the agent proceeds to offer personalized exercise recommendations, where the system provides up to five exercise options based on the curated list of exercises from the predefined list of exercises. More specifically, we used the list of exercises from Agapie *et al.* [5] that contains 112 common exercises that were curated by the expert exercise planners. The list contains the name of the exercise, as well as its alternative names (if any), intensity, laypeople description (*e.g.*, definition, how to perform), and the muscles involved/exercise type. As the list is stored and loaded in CSV format, it can be easily expanded by altering with external exercise databases in the future if needed.

The recommended exercises are displayed on the dashboard with a brief description, which summarizes the definition of the exercise and the reasoning behind the recommendation (Figure 1-Ⓑ). For users seeking more comprehensive information about a particular exercise, a 'more' button is provided where the users may click to retrieve additional details of the exercise from the attached database. Then, users are asked to select their desired exercises by either typing the name of the exercise(s) into the chat panel in a free form or clicking on them on the dashboard; if they wish to explore additional exercise options, they are also allowed to simply ask a request to the CA, which will result in refreshing the recommendations.

4.1.3 Generating a personalized exercise plan. After the user finalizes the exercise types, PlanFitting generates and outlines an exercise plan, with its structured format displayed on the dashboard (Figure 1-Ⓒ).

Format of the plan. Our interview study suggests that prescribing exercise broadly (*e.g.*, specifying a weekly amount) could burden users with scheduling and possibly lower motivation. Thus, to better contextualize the exercise plan within the user's availabilities, PlanFitting offers each exercise plan in an *implementation intention* [31] format, a grounded strategy rooted in behavioral psychology that aligns the user's intentions with specific events, hence offering a structured format in well-established `IF-THEN` statements. (*i.e.*, "`IF` *{availability (time or situation)},* `THEN` *do {exercise type} for {amount} at {intensity}*") In addition, the agent offers a *coping plan* for each plan, which equips users with an alternative plan to follow when the original plan cannot be executed due to the obstacles that may happen. (*i.e.*, "`IF` *{obstacle},* `THEN` *{alternative}*")

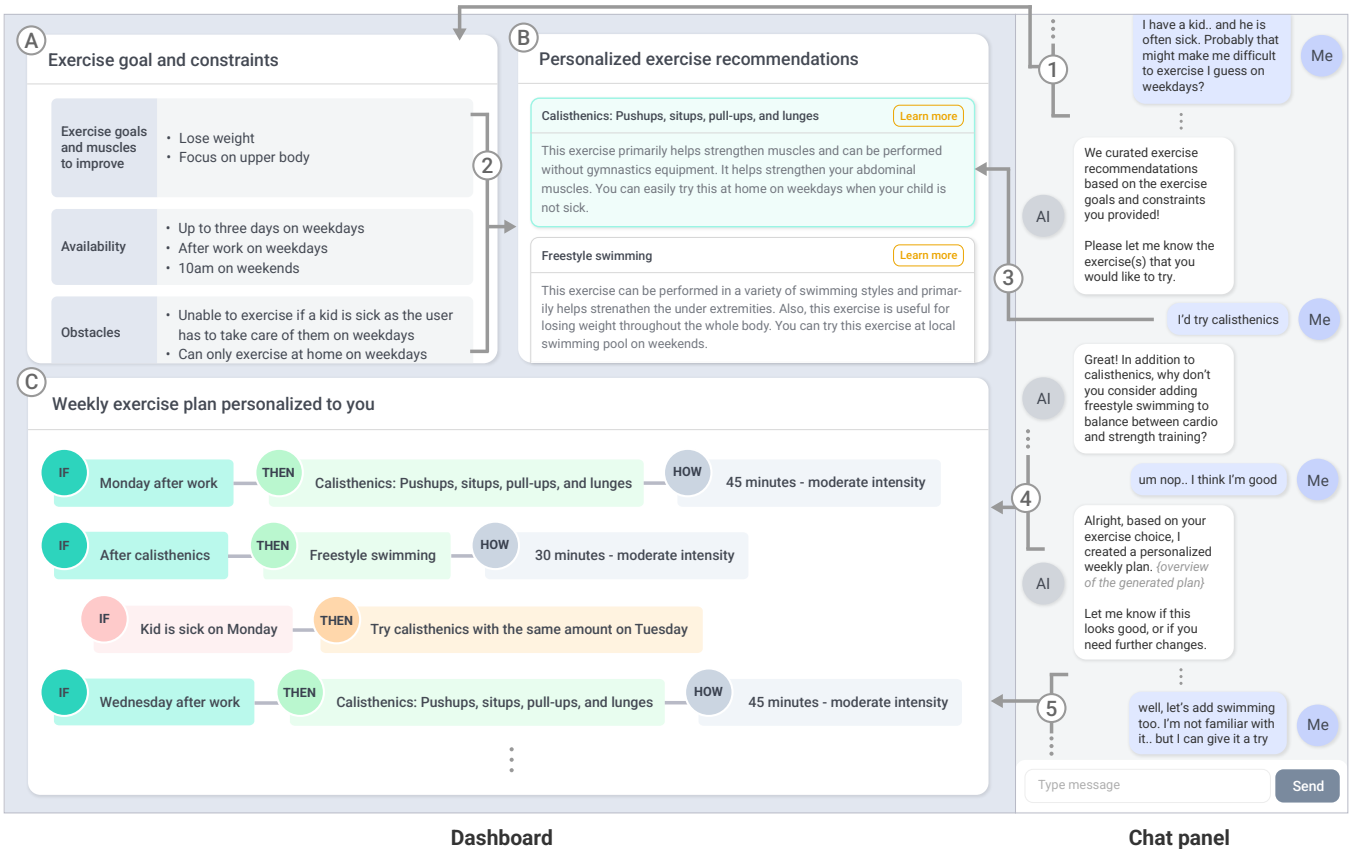


Figure 1: Key screen and interaction flow of PlanFitting. ① Once the user describes the goal of the exercise and their own constraints in a natural language on the chat panel, they are parsed and synchronized with the dashboard. ② Based on the collected information, PlanFitting recommends exercises and ③ the user can provide the exercise type(s) they want to include. Once the user finalizes exercise types, ④ the agent returns a weekly exercise plan, where the user can ⑤ continuously iterate on the plan through natural language.

Grounding a plan to global exercise guidelines. To earn rigor for the generated plans, the agent applies a common set of guidelines that we elicited from the recommendations offered by the universally recognized exercises guidelines (*i.e.*, ACSM [3, 28], U.S. Department of Health and Human Services [58, 65]), as well as their previous application to the technology-mediated exercise planning [5]:

First, the agent is instructed to allocate exercises totaling more than 150 minutes per week [3, 5, 28, 58, 65]. To comply with the guidelines, it also accounts for vigorous-intensity exercises by doubling their allocated time when calculating the total exercise duration [5, 58, 65]. In addition, to balance between cardio and strength training [3, 5, 28], if the user had initially chosen exercises of either type only, the agent asks users to consider incorporating both types of exercise. Lastly, the agent puts a minimum of a one-day rest period between exercise sessions, if the user constraints allow, to prevent any potential negative effects of consecutive days of exercising the same or adjacent muscle group [3, 5, 28].

4.1.4 Revisiting and refining the exercise plan. Following the initial planning phase, PlanFitting is designed to allow for iteration of the plans by inquiring users about their satisfaction with the existing plan, when the user returns to the system. More specifically, PlanFitting is instructed to first ask the user whether they followed the previous week's exercise plan and whether they were satisfied with it. If the user is satisfied with their plan, the agent asks if they are willing to extend the allotted time to adhere to the progression principle (*i.e.*, gradually increase the engagement in exercise) of the exercise [3]. Otherwise, if the user indicates dissatisfaction, it solicits feedback on the specific aspects that require revision, facilitating an iterative approach to refining the plan. As such, the agent enables ongoing, open-ended planning, conducive to continuous improvement based on user input.

In summary, the interaction flow between the user and the CA is structured to facilitate user engagement, provide exercise recommendations, and enable the creation of personalized exercise plans that adhere to recognized exercise guidelines, while allowing for the iteration of generated plans.

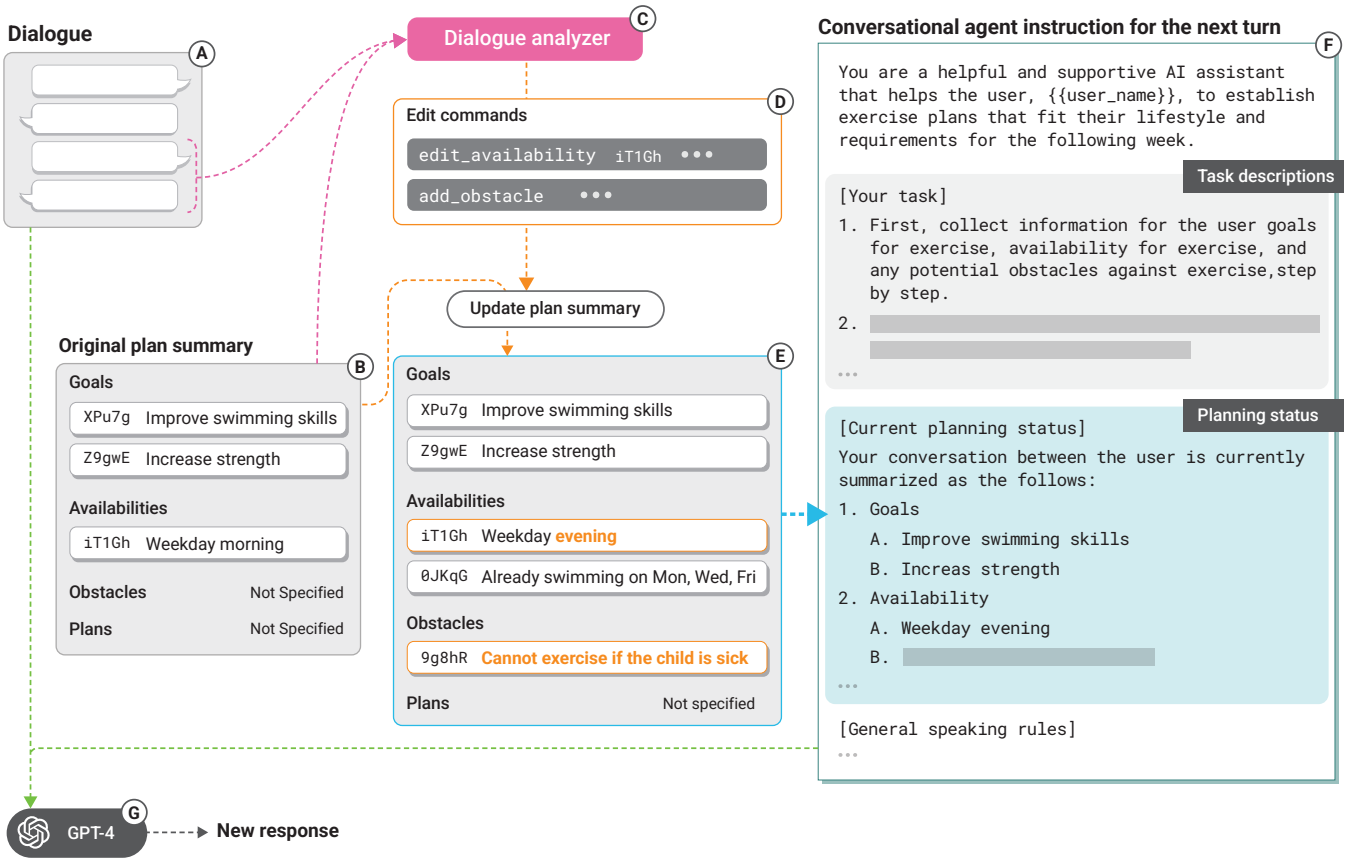


Figure 2: Illustration of how the PlanFitting computes and returns the next dialogue of the conversational agent and updates the dashboard based on the current dialogues

4.2 Conversational Pipeline Design

Figure 2 illustrates the pipeline of PlanFitting's CA system. PlanFitting's CA is driven by two LLM components: a **response generator** (Figure 2-Ⓖ) and the **dialogue analyzer** (Figure 2-Ⓒ). The response generator generates the agent's response based on a global instruction (Figure 2-Ⓕ) and the current dialogue (Figure 2-Ⓐ). The user's constraints and generated plans are maintained in a data structure called "plan summary" (Figure 2-Ⓑ), which maintains the current status while providing information to be displayed on the dashboard UI.

Plan summary update. Inspired by memory management techniques from the NLP discipline (*e.g.*, [8]), we designed the dialogue analyzer to generate edit commands that modify the previous state of the plan summary. The dialogue analyzer receives the latest turn pair (*i.e.*, the CA's message and the user's response; Figure 2-Ⓐ) and the plan summary of the previous cycle (Figure 2-Ⓑ) as inputs and generates a list of edit commands (*e.g.*, add, update, and remove; Figure 2-Ⓓ) that reflect the changes caused by the new messages. Then, the system applies the edit commands to the plan summary and generates a new plan summary (Figure 2-Ⓔ). The CA updates the plan summary every time before it generates and returns a response to the user. The base prompt of the dialogue analyzer is in Appendix A.1.

Conversation. Once the plan summary is updated, an instruction prompt (Figure 2-Ⓕ) is formulated and fed into the response generator. The instruction includes the task descriptions on how to carry on the conversation (Figure 2-Ⓕ, Task descriptions), and the current plan summary to inform the model with which constraints are missing, thus what needs to be asked in the following dialogues (Figure 2-Ⓕ, Planning status).

When defining tasks for exercise type recommendation and generating plans, we established rules to append XML data to the message so that the system can systematically parse responses and integrate them into the user interface. For example, we specified the message rules for creating the plan as follows:

```
Using the exercise types that the user se-
lected, plan for and return the user's ex-
ercise plan in the implementation intention
format
...
Each implementation intention rule should
be accompanied by corresponding coping plans
```

```
that can be plan B when the user fails to ad-
here to meet the main rules. It should assume
the failure of each of the user's availabili-
ties due to the obstacles the user mentioned
...
Each exercise/coping plan should be described
in an IF-THEN format along with AMOUNT in-
side
...
(Example:
<If>Monday after work</If>
<Then>
  <Exercise>Running</Exercise>
  <Amount>60 minutes - moderate intensity</Amount>
</Then>
<If>After running</If>
<Then>
  <Exercise>Pilates</Exercise>
  <Amount>30 minutes - vigorous intensity</Amount>
</Then>
<If>Too sleepy after work on Monday</If>
<Then>
  <CopingPlan>Do the same exercises on Tues-
day</CopingPlan>
</Then>)
```

To implement exercise recommendations, we employed function calling [62] to extract exercise-related keywords from the user dialogue and enable structured data retrieval. Coupled with function calls, we used cosine similarity to compute the semantic closeness between the user's input and each predefined exercise description. We believed leveraging cosine similarity to be particularly suitable for this task, as it captures semantic similarity between vectors while making it robust to variation in user phrasing. To support this, we generated vector representations (*i.e.*, word embeddings) using a pre-trained sentence embedding model (`text-embedding-ada-002`), which encodes both user input and exercise metadata (*i.e.*, name, description, involved muscles) into a shared semantic space. By computing cosine similarity between these embeddings, the system identifies exercises that are semantically aligned with the user's intentions, even when those are expressed in informal or diverse language.

More specifically, we first embedded the title and description of each exercise from our prepared list and saved them as embedding vectors. Once the user finishes providing their constraints and an LLM detects if the exercise recommendation is needed, a function calling that takes the goal and obstacles as input and returns the recommended exercises is triggered. Here, the function is programmed to embed the detected exercise-related keywords of users as an embedding, which is then compared to each embedding of each exercise from the list to calculate the cosine similarity and return the types of exercise that have the top 5 cosine similarities to the user in a JSON format. Then, similar to how the system does for generating the exercise plan, PlanFitting formats the output JSON to the XML format through Regex postprocessing, which is then populated in the dashboard (Figure 2-Ⓑ).

4.3 Implementation

PlanFitting system consists of two components: (1) a conversational UI and (2) a backend server, where the user interacts with the interface, whose chat is computed to return the response from the backend server.

The conversational UI is built as a web-based application on a JavaScript-based framework (SvelteKit). For the backend, we employed a Python Flask API server that takes the user's name and chat message as inputs and generates the subsequent message along with detected metadata–including exercise goals, availability, constraints, and recommended/selected exercise types. To run LLM computations in our conversational pipeline components, the system uses OpenAI [61] GPT-4 model with the following parameters: $temperature = 0.5$, $top\_p = 1$, $frequency\_penalty = 0$, and $presence\_penalty = 0$.

5 USER STUDY

To gain a comprehensive understanding of the use of PlanFitting, we conducted a user study where participants interacted with PlanFitting to set up their exercise plans with the CA based on their own goals and constraints. In addition, the adherence to the guidelines and rigor of these plans was later assessed through intrinsic evaluation and expert evaluation, respectively. The study protocol was approved by our institutional review board.

5.1 Participants

We posted our study recruitment to a local online community platform and the corporate bulletin board, where we required participants to be (1) aged over 19, (2) motivated to do regular exercise, and (3) not currently doing exercise under the specific plan advised by planners to avoid any conflict, (4) who can participate in an in-person lab study. As a result, we recruited 18 participants (P1–P18; 11 females and 7 males) who were aged between 19 and 54 ($M = 33.2$). Of all, six were full/part-time employees by the time they were participating in our study, six were college students, one was a retiree, and five responded that they were either stay-at-home parents or unemployed. We compensated 50,000 KRW (approximately 35 USD) as a gift card for their participation.

5.2 Study Procedure & Tasks

To explore how the participants create and refine their exercise plans with PlanFitting, we structured the user study in the following phases: (1) initial planning, (2) iteration, and (3) debriefing. Throughout the planning, each participant was asked to think aloud in order for us to better surface their lively experience interacting with the CA.

Initial planning. The initial phase involved participants being guided through the process of configuring their exercise plans with the assistance of PlanFitting. Participants were asked to interact with the agent to articulate and input their specific exercise goals, availabilities, and any potential obstacles. At the same time, they were also encouraged to freely ask questions to the agent and iterate on their plans until they were satisfied. As such, we aimed to mirror the process of tailoring exercise plans to individual constraints based on the overall guidance of the PlanFitting system.

Iteration. After setting up their weekly exercise plan initially, participants were instructed to move on to the second phase. In this phase, they were asked to imagine themselves in the upcoming week, having completed their exercises successfully, and to also consider scenarios that may have hindered their progress in the previous weeks. To assist them in this process, we presented example scenarios for their reference (*e.g.*, *"I intended to swim last week, but I'd rather avoid such location-dependent activities due to the hassle of making reservations"*). In cases where they had nothing to change, they could engage with the system as if they were satisfied with their plan.

Once they had formulated their scenarios, participants were encouraged to use PlanFitting to review and fine-tune their exercise plans over a designated time frame. They were asked to freely describe adjustments to the agent that they would want to make, such as exercise availabilities, types, and amounts.

Debriefing. During the final debriefing phase, we conducted a survey and a semi-structured interview with each participant to gather their feedback, insights, and reflections on both the planning process and their interactions with PlanFitting.

The survey was designed to assess their subjective evaluation of how personalized and actionable the plan they created with PlanFitting is, as well as their degree of acceptance and adoption of the PlanFitting system. To evaluate the level of personalization and actionability, we measured *follow* and *fit* for personalization, and *specificity*, *encouragement*, *vocabulary*, and *accuracy* for actionability on a 7-point Likert scale, following the rubric from the prior literature [5] that evaluated the quality of the plan. For evaluating the acceptance and adoption of PlanFitting, we used the Technology Acceptance Model (TAM) scale [73]. The whole procedure was conducted on the user's private screen to reduce bias.

Then, we conducted an interview, where we inquired about the overall usability, their feedback on the iteration process with the agent, the quality of the generated plans, and the potential future enhancements. The overall procedure took approximately 1 hour for each participant.

Intrinsic evaluation. The research team evaluated how well each participant's plan followed global exercise guidelines after both the initial planning and iteration phases. Two researchers manually read through each plan, discussed, and reached a consensus on whether each plan met the requirements of the global guidelines, as detailed in Section 4.1.3, including (1) amount (*i.e.*, whether the total exercise time of the weekly plan exceeds 150 minutes, while counting vigorous activity double), (2) balance (*i.e.*, whether both cardio and strength training are included in the plan), and (3) resting (*i.e.*, whether one or more rest day(s) between exercise days are included).

Expert evaluation. To assess the generated plans from the perspective of experts, we recruited three expert planners (E1 – E3; one male and two females) from an in-house clinic of the corporation. The experts were nationally licensed physical therapists aged between 28 and 39 (M = 31.3), and had an average of 7 years in professional exercise planning (SD = 4.6). We asked the experts to evaluate plans from the initial exercise planning phase both quantitatively and qualitatively, where each expert was randomly assigned six plans and asked to evaluate them.

Specifically, the experts holistically reviewed the plans as well as the constraints and conversation history, with private information masked. For each plan, they filled out our evaluation form that consists of a 7-point Likert scale of four items from the FITT principles [20]—*frequency* (*i.e.*, how often the exercises in the plan are), *intensity* (*i.e.*, how intense the exercises consisting of the plan are), *time* (*i.e.*, duration of the exercises consisting of the plan), and *type* (*i.e.*, composition of the types of exercise consisting of the plan)—a recognized and empirically validated framework consisting of salient factors in exercise plan design and assessment (1: highly unsatisfactory, 7: highly satisfactory). For each item, we also included an open-ended field asking for the rationales for the assessment.

5.3 Analysis

Similar to what we did for our formative study, we used a thematic analysis to code (1) participants' responses and (2) qualitative responses from expert evaluations. The two authors of this work first read and gained a sense of the raw responses independently, and each author identified emerging themes from the responses. Then, they teamed up to discuss and compare the themes during the regular meetings until they reached a consensus.

Additionally, we analyzed the interaction logs to understand the interaction between the participants and the CA. The two authors first individually reviewed the logs, linking each user action to the user-defined constraints (*i.e.*, goal, availability, obstacle) and exercise type. Based on this review, we initially classified each action as `add`, `edit`, or `remove`, denoting whether it aimed to introduce a new entity, modify an existing one, or delete one. The research team then met three times to conduct a bottom-up thematic analysis to discuss and consolidate the emerging categories that could be characterized as distinct actions. Through this process, we additionally identified and defined new action types such as *amount* (*i.e.*, adjusting the exercise amount), *question* (*i.e.*, asking the agent questions), and *querying exercise list* (*i.e.*, requesting exercise recommendations based on user-specified constraints). These actions were then organized in a sequence for each participant.

6 RESULTS

In this section, we report the results of our study in three parts: (1) collected constraints and interaction patterns with the conversational agent, (2) user evaluation of the agent and its crafted plans, and (3) quality of the generated plans.

6.1 Collected Constraints and Interaction Patterns

As shown in Table 1, through interacting with the agent, participants provided a wide range of constraints related to their lifestyle for crafting a personalized plan. On average, participants shared 2.28 exercise goals (SD = 1.04), 1.72 availabilities (SD = 0.80), and anticipated 1.33 potential obstacles (SD = 0.88). Some common goals that the participants described include weight loss (N = 11), recovering daily energy (N = 8), and maintaining/improving muscular strength (N = 5). For availability, only 5 participants described

Table 1: Exercise goals, availabilities, and potential obstacles that PlanFitting surfaced from the participants during the initial planning phase

ID	Goal	Availability	Potential obstacle
P1	• Weight loss • Recover energy	• Weekdays at night after 6 pm • Weekends in the morning	• Do not wanna do exercises that heavily affect knees • Company dinner or other appointments
P2	• Maintain muscular strength • Be more energetic in daily life • Weight loss • Maintain daily health • Cardio	• After waking up • If it fails, exercise afternoon or at night instead • Light exercise after lunch	• Light exercise at night • Hard to exercise on the day after drinking • Sudden schedules afternoon • Sudden schedules at night
P3	• Recover basic energy	• After school	• Difficult to exercise after heavy drinking
P4	• Weight loss • Overcome exercise shortage since pandemic	• Thu–Sun after 7 pm	• Don't want to exercise on rainy days
P5	• Improve muscular strength • Fix posture	• Everyday in the morning	• Want to exercise without equipment • Not familiar with exercise
P6	• Weight loss • Improve shoulder muscles • Relieve wrist pain	• Everyday in the morning except for late night	• Diagnosed with right shoulder subluxation
P7	• Recover energy • Weight loss • Improve muscles	• Weekdays in the morning & at night	• Kids' day off from school or appointment • Kids/husband come back home early
P8	• Weight loss • Recover energy • Relieve stress • Get hobbies	• Weekdays in the morning • Weekdays afternoon • Weekends at any time	• Difficult to exercise after drinking or sleeping late • Postpone exercise if there is a schedule with others
P9	• Improve swimming skills • Improve muscular strength	• Weekdays in the morning • Unable to exercise on Mon–Fri as already doing swimming	• Difficult to exercise if a kid is sick
P10	• Weight loss • Recover energy	• Weekdays after school at night • Weekends afternoon • Tuesday afternoon–night	• Sleepy after school
P11	• Weight loss • Cardio	• After dinner	• Location constraint
P12	• Weight loss	• Everyday in the morning & at night	• Diagnosed with back disc
P13	• Weight loss	• Three times per week in the morning (9–12 am)	• Prefer indoor exercise • Diagnosed with peripheral edema
P14	• Weight increase • Recover energy	• Everyday after 7 pm except for Sat	• Want to avoid excessively using the right index finger
P15	• Improve arm muscles • Want to make waist look thinner	• Weekdays at night • Weekends 10–12 am	• Weekday night party • Wish to exercise three times per week
P16	• Weight loss • Relieve waist pain • Get broad shoulders	• Tue–Thu after school • Fri & Sat before work • Sun & Mon at anytime	N/A (provided no obstacle)
P17	• Improve golf–backswing skills	• Mon at anytime • Thu & Fri at night	• Economical exercises
P18	• Recover energy • Improve muscles • Relieve back pain	• After work • Weekends afternoon	N/A (provided no obstacle)

their availability in the exact time format (*e.g.*, after 7 pm); the others described all of their availabilities freely in a descriptive form (*e.g.*, after school). Lastly, participants described their potential obstacles in a highly personalized expression by drawing connections to various aspects of their own lifestyles and circumstances, such as heavy drinking (P2, P3), kid's schedule (P7, P9), and party (P15).

Figure 3 illustrates how these constraints were provided to the conversational agent and modified for each participant across the

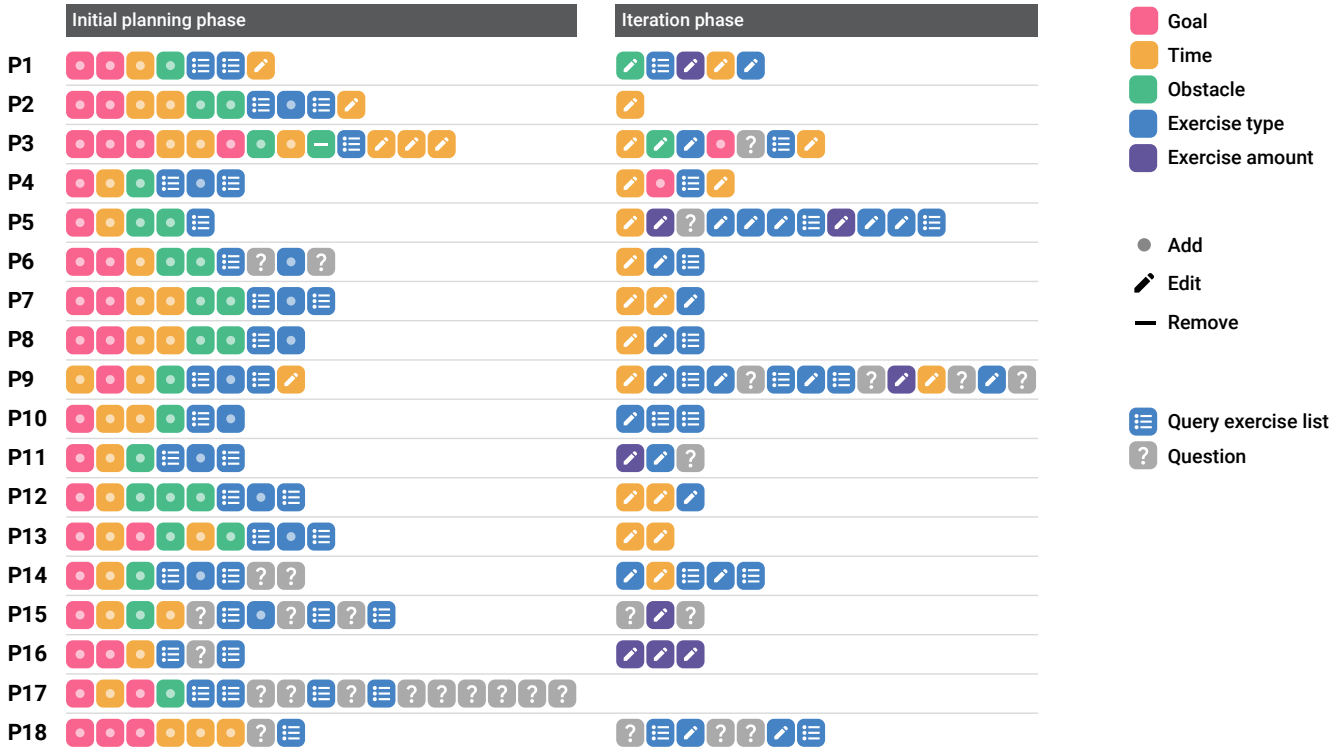


Figure 3: Sequence of how the participants interacted with the conversational agent to tailor their exercise plan

two study phases. In the early stage, participants generally followed the sequence of information that the conversational agent's dialogue analyzer was instructed to collect. As the interaction progressed, they iterated on their constraints in individual ways through flexible conversational interaction. Ten participants (56%) also asked questions (Figure 3; gray rectangles) to PlanFitting about exercise and other relevant topics.

6.2 User Evaluation

Table 2a illustrates the technology acceptance scales, and Table 2b and 2c illustrate the distribution of user evaluation assessing the quality of plans guided by PlanFitting (*i.e.*, personalization, actionability), respectively. Below, we describe the participants' quantitative evaluation in detail, along with their qualitative feedback, offering insights into the role of the agent in creating quality plans and the user perception of CA-supported exercise planning.

6.2.1 Perceptions of the use of conversational agent. Participants indicated a positive inclination towards interacting with the agent, where the perceived usefulness received a rating of 5.43 on average ($SD = 0.99$). Similarly, participants found PlanFitting to be easy to use, with an average of 6.00 ($SD = 1.12$). As for the intention to continue using it, participants responded with an average rating of 5.52 ($SD = 1.26$).

From the interview, we could uncover factors that contributed to the participants' intention to keep using it. First, free-flowing and flexible conversations enabled by the LLM-driven conversational agent allowed participants to freely initiate actions (*e.g.*, introducing additional constraints) whenever they naturally came to mind, not only limited to when prompted by the agent's questions. This flexibility enhanced their perception of the agent's utility in the planning process: *"Even if I suddenly went back to a previous question or said something else, the system seamlessly continued the conversation which made the chatting more convenient."* (P3)

Second, displaying the conversational history in a dashboard alongside the chat panel allowed participants to easily track the constraints they had specified. This visibility helped them adjust plans efficiently without needing to manually navigate through the entire chat history. With a clear overview, participants could stay aware of all constraints and correspondingly updated plans without having to read through lengthy dialogues, increasing their confidence in using the agent continuously: *"The dashboard neatly organizes and updates the information every time I entered constraints, which I find very convenient. Often times when I plan things like this, I have to make separate notes on my phone, right? Now I can just provide it to the agent, and it automatically organizes it for me (...) I consider this as a very useful component."* (P12)

At the same time, participants also anticipated the future integration of additional contextual information that the agent could collect and consider during the personalization process. For instance, some participants suggested that incorporating context-aware features, such as providing exercise recommendations based on their current location and weather conditions, could significantly enhance the system's utility. Additionally, soliciting more detailed constraints

Table 2: Mean participant ratings with standard deviation (7-point scale) across evaluation categories.

(a) Technology acceptance			(b) Personalization		(c) Actionability			
Usefulness	***Ease of use***	***Intention to use***	***Follow***	***Fit***	***Specificity***	***Encouragement***	***Vocabulary***	***Accuracy***
5.43 ± 0.99	6.00 ± 1.12	5.52 ± 1.26	5.83 ± 0.99	6.00 ± 0.77	5.06 ± 1.51	5.56 ± 1.34	6.19 ± 1.11	5.72 ± 1.23

from participants, such as whether they have children (P13) or specific muscle areas requiring rehabilitation (P15), was identified as a future enhancement that would further enhance the usefulness of the agent.

6.2.2 Personalization. Overall, participants found that interacting with PlanFitting helped them to generate personalized plans. They reported the plans to fit their personal lifestyle (M = 6.00, SD = 0.77), and that they were generally likely to follow the plans (M = 5.83, SD = 0.99), as evaluated on a 7-point Likert scale (1: strongly disagree, 7: strongly agree).

Participants reported that the expressivity of the conversational agent when guiding them, as well as its understandability of user requests in real-time, greatly contributed to successfully tailoring exercise plans according to their preferences and constraints. In contrast to constrained expert-client communication settings, PlanFitting allowed participants to make unlimited requests easily through natural language. This served as the flexibility to personalize their plans as extensively as they desired: "*It was refreshing to have schedules tailored to my personal time and listen to my request. I was really surprised to see that it could do that well.*" (P15) Observing the output plans where the agent successfully incorporated these personal requests and constraints in real-time, participants found the planning process with PlanFitting to be highly personalized: "*Once I requested, it extended the duration of each exercise session by 15 minutes.*" (P15); "*I said I wanted to do simple, sweat-free, and noiseless exercises at home. Tailoring my plan using this, it was great to see that my preferences and conditions were reflected exactly in the plan that looks easy to follow.*" (P10)

Not limited to the initial generation of plans, participants also highlighted the agent's potential to let users reiterate their existing plans and constraints over time. For instance, when unforeseen changes occur, which could make it difficult to adhere to the original plans, PlanFitting's capability of facilitating iterative planning would greatly help them quickly adapt to unexpected changes: "*It is really nice having the option to easily modify the existing exercise plan when a new goal arises (...) For example, if I suddenly injure my leg and need rehabilitation, I'm sure it would also be well-reflected in my plan.*" (P16); *"If there's a change in my availability, being able to make adjustments instantly like this, I believe I would use it frequently."* (P14) With such support that enables users to freely iterate on their plans, participants came up with various potential use cases of PlanFitting, such as finding and engaging in lightweight exercises that can be done on the go or when they suddenly have some free time: *"Let's assume that I want to utilize some spare moments, for example, when I finish lunch early and have about 20-30 minutes left. Then I could easily use this system in my workplace to use those spare moments."* (P5)

6.2.3 Actionability. As in Table 2c, participants were also positive about the actionability of the plans they created with PlanFitting, which were generally received to be specific with enough details to act upon (M = 5.06, SD = 1.51). Participants also found the agent's presentation of the plan and its accompanying information encouraging (M = 5.56, SD = 1.34), described with straightforward vocabulary (M = 6.19, SD = 1.11), and accurate (M = 5.72, SD = 1.23), responded on a 7-point Likert scale (1: strongly disagree, 7: strongly agree).

Our qualitative analysis revealed that participants found the `IF-THEN` implementation intention format presented by the agent practical and adaptable, especially for individuals with fluctuating schedules. Avoiding vague timing instructions (*e.g.*, 3 times per week) or rigid time constraints (*e.g.*, 7 pm) while contextualizing the plan to the user-described situations, the plan was perceived as realistic and easy to remember, making participants find it to be more actionable: "*I think it's better when it tells you to do some exercise based on the situation like this. Honestly, sticking to a set time isn't always easy to follow through with, in reality.*" (P6)

Furthermore, participants reported the plans guided by the agent to be well-adhering to the specific constraints they provided: *"For every information I added to the chat, the system successfully reflected those to my exercise plans."* (P3) The plans were also reported to be presented in sufficient detail to follow by specifying the exercise type and amount, which was viewed as clear and easy to follow: "*What surprised me was how it instructed me on what to do on each day, like there was a clear outline. I liked that it was so specific. I tend to prefer clear instructions (...) Nowadays, there are just too many choices, and I tend to dislike making decisions. So, having such clear instructions made me appreciate why I should use this and why I rated it highly.*" (P4) With such specificity of the plans, participants noted that the generated plans are systematic and presented in an actionable format: *"I felt like I could systematically handle various types of exercises a bit better. It gave me a feeling of being well-grounded."* (P6)

Participants also highlighted that providing coping plans for each exercise plan contributed further to its actionability. They expected that, even when facing obstacles that might lead them to skip an exercise, these coping plans would serve as clear guidance to make up for the exercise: *"If I find myself unable to do my exercise and I'm debating whether to skip it for the day, seeing this alternative [coping plan] might make me think, 'Well, if I can't follow the original plan, I might as well do the alternative one today,' and it would induce to start exercising anyway."* (P18)

6.3 Evaluation of the Quality of Generated Plans

6.3.1 Intrinsic evaluation. From our intrinsic evaluation, we found that PlanFitting effectively guided participants in aligning their plans with the global guidelines, as outlined in Section 4.1.3:

(i) Amount. PlanFitting successfully met the amount guideline while calibrating amounts across individual sessions. Of the initial plans, 15 of 18 (83%) reached the required weekly total; two participants received less as they mentioned they were already personally engaging in other activities outside the planning, and one participant manually asked to exclude a session. In the iteration phase, all but three participants (including two from the initial phase) finalized plans that satisfied the amount guideline.

(ii) Balance. PlanFitting allows users to first choose their preferred exercise types, while suggesting adding complementary types if the exercise types from only one category (cardio or strength) were chosen. In the initial planning phase, nine participants selected either cardio or strength only, prompting PlanFitting to recommend adding the missing type, where eight of the nine accepted to create a balanced routine. During the iteration phase, one participant provided an injury-related constraint, leading PlanFitting to remove certain exercises and omit one type.

(iii) Resting. Of all, 14 plans (78%) in the initial phase met the guideline for ensuring a rest day. In the exceptional four cases, the system could not include this gap as participants manually indicated the days of the week for exercise. In the iteration phase, three more participants requested schedule changes based on their scenario which necessitated consecutive exercises; for all the others who did not add such inevitable constraints, the iterated plan satisfied the guideline.

6.3.2 Expert evaluation. Expert planners generally found the plans generated by PlanFitting to be well-adhering the FITT principle—how adequately the *frequency*, *intensity*, *time*, and *type* of exercises were formulated (see Table 3). Below, we describe the assessment and feedback we gained from the experts and the potential room for further enhancing the plans.

(i) Frequency. Experts generally rated the exercise frequency of the plans as well-defined, averaging 5.67 on a 7-point Likert scale ($SD = 1.53$). Qualitative feedback highlighted the agent's success in accommodating both the 150 minutes per week guideline and individual preferences, particularly its approach of evenly distributing exercise throughout the week while incorporating rest days: "*It's highly commendable to reflect the exercise guideline by scheduling exercise with the assigned time for at least 3 times a week and incorporating the concept of rest on the day after exercise.*" (E1) However, experts also suggested future improvements, including prompting the agent to adjust frequency based on the number of exercise types selected, potentially increasing frequency for plans corresponding to the number of exercises the participants wish to do (E3: *"Given the four different exercise types (that the participant mentioned they wished to do), it may make sense to increase the exercise frequency from the current four times a week to five or six."*) and decreasing frequency for plans with similar exercises to avoid muscle fatigue and injury risk (E2: *"The plan consists of 7 days of exercise sessions that target the abdomen and lower body, which could potentially lead to muscle fatigue. It's essential to reduce the frequency."*)

(ii) Intensity. The experts rated the exercise plan's intensity with a general favorability with room for improvement, with an average score of 4.28 ($SD = 1.32$). They praised the system for preventing intensity-related issues through coping plans based on participant-reported obstacles, such as advising participants with back pain to stop exercising and consult specialists if needed: *"I found cautionary comments for the patients with back pain to be great, along with the appropriate intensity of exercise offered."* (E2). However, experts suggested enhancements to the system's guidance on intensity. The agent currently recommends increasing the amount of exercise as a progression measure if users are satisfied with previous plans. On top of the time, E1 suggested that the intensity of the plans can also be used as a measure for the progression: *"In terms of the intensity of this plan, I consider it appropriate. Given that the participant is healthy, I also recommend the user start with moderate intensity and gradually progress to higher intensity."* Similarly, the agent uses predefined intensity information in our predefined exercise list to guide recommendations, but based on the user needs like weight loss or muscle strength improvement, E1 suggested customization to include high-intensity exercises corresponding to participants' individual goals: *"To achieve weight loss, I believe it is necessary to include high-intensity aerobic exercises that have a higher level of intensity."*

(iii) Time. As detailed in Section 6.3.1, the agent effectively generated exercise plans that comply with the ACSM guidelines for exercise time. The evaluation of these plans particularly praised the adherence to the time component, with a rating of 5.06 ($SD = 1.80$). Experts expressed satisfaction with the planning, and offered recommendations to improve flexibility. For example, if a user is unable to commit to a 30-minute exercise, E1 suggested it could be further broken into shorter sessions (*e.g.*, three 10-minute sessions) for flexible planning: *"I think the amount of time has been planned well. If the client is unable to commit to a 30-minute exercise, you can also advise them to break it down into three 10-minute sessions."* (E1) Additionally, there is potential to enhance PlanFitting by operationalizing exercise time not just in weekly totals but also in per-session durations. While the system meets the ACSM guidelines for total weekly duration, experts identified areas for improvement in individual sessions. For example, if a user has limited time for exercise, the agent currently generates long, higher-intensity sessions to meet the guidelines within fewer available days. However, planners cautioned that such prolonged, intense sessions could lead to overexertion, advising against these exceptional cases: *"For the case of high-intensity exercises, prescribing a 50-minute session of strength training is excessive for the participants."* (E2)

Table 3: Mean expert ratings with standard deviation (7-point scale) across FITT principles.

Frequency	Intensity	Time	Type
5.67 ± 1.53	4.28 ± 1.32	5.06 ± 1.80	3.89 ± 1.45

(iv) Type. The exercise types within the plans received a slightly below satisfactory rating of 3.89 (SD = 1.45), emphasizing the need for enhanced tailoring. Expert feedback highlighted key areas for improvement, particularly in guiding users to balance cardio and strength exercises; while the agent already encouraged users to include at least one exercise of the opposing type, it did not enforce equal distribution, resulting in some plans being heavily skewed or sometimes omitting one category. To address this, E2 suggested adopting a more assertive tone when presenting recommendations to ensure balanced planning: *"Only the exercises the user wanted to do were included. However, as this is an interaction where AI sets exercise goals together with the participant, 'necessary exercises' should also be guided."* Also, experts identified inaccuracies when specific muscle groups were not explicitly mentioned in participants' goals—such as the agent relying on cosine similarity between "golf" and exercise descriptions, which led to overlooking beneficial strength and flexibility routines. E3 pointed out this limitation, suggesting that enhancing PlanFitting to infer relevant muscle groups, even when not explicitly stated, could improve the accuracy of exercise recommendations: *"Other exercises that could enhance golf performance were not adequately suggested (...) recommendations for improving golf backswings should include exercises that enhance flexibility, core strength, and lower body strength."*

7 DISCUSSION

In this section, we discuss lessons learned from designing and implementing a CA for personalized exercise planning, as well as its evaluation from our user study.

7.1 Leveraging LLM-driven CAs for Exercise Planning

Our work proposed leveraging LLM-driven CAs to create personalized exercise plans that account for individual constraints, while aligning the plans with global guidelines. Instead of relying solely on simple LLM generation based on the knowledge base of generic models—which may be prone to hallucination [35] and lack of the output's alignment with the real-world practice and guidelines [53], we developed a pipeline that integrates expert-verified exercise lists and guidelines to inform the generated plans in a way better aligned with real-world practices, which received positive feedback in expert evaluations. Additionally, visualizing the current planning status on a dashboard helped participants better keep track of their plans, without losing the context while engaging in the back-to-back conversation.

Building on these, the free-form conversations carried by LLM-driven CAs enabled participants to provide their exercise-related constraints intuitively and flexibly, resulting in the system identifying diverse and unique constraints from participants (*c.f.*, Table 1). Additionally, the conversational interaction allowed the exchange of questions and reiterations of the plans (*c.f.*, Figure 1), seamlessly interleaved in the user interface. This was shown to be effective during the iteration phase of the study, where PlanFitting successfully adjusted the plans per user requests. Observing the system reflect their requested edits in the plan, participants expressed intention to use PlanFitting in the long term and frequently throughout their exercise journey. Our work suggests that LLM-driven conversational interaction could successfully simulate natural interactions in exercise planning settings, while demonstrating opportunities for long-term engagement with an exercise assistant agent.

In this process, unlike typical open-domain conversations where most LLM-driven CAs operate, PlanFitting needed to reliably adhere to user-defined constraints and create exercise plans grounded in established guidelines. To achieve this, we employed several design choices to enhance compliance with our design goals. First, we incorporated two distinct agent routines: one dedicated to generating conversational dialogue and another for transforming user dialogue into formatted data, used for input summaries and exercise plan generation. Having two routines dedicated to conversation and analysis respectively, our CA could attain reliability in both tasks. Second, to enable the agent to reference external exercise knowledge, we implemented a retrieval-augmented generation technique to integrate an existing exercise database, allowing for more evidence-based planning while preserving the flexibility of agent-driven conversations. Additionally, this approach would allow PlanFitting to easily tailor its focus on specific exercise environments (*e.g.*, bodybuilding) or organizational settings where clients have access to only a restricted set of exercises, simply by replacing or modifying the exercise database.

7.2 Incorporating Nuanced Perspectives of Domain Experts

Although PlanFitting generally complied with the exercise guidelines as intended, evaluation from the expert also revealed the future enhancements for some components of the crafted plans (*i.e.*, exercise intensity, types), suggesting edits that they would have applied to the plans based on their own hands-on experiences (*e.g.*, recommending a certain exercise intensity for achieving specific exercise goals). This points out that, although PlanFitting is reported to successfully take into account various individualized factors and exercise guidelines during the exercise planning process, human expertise may still contribute to enhancing the quality and effectiveness of the plans. Since such edits and potential contributions may be grounded upon the experts' tacit knowledge from the lessons they learned over time, it is not trivial to formalize such knowledge into global guidelines and reflect them to the agent's instruction.

A promising way to address this gap is through multi-agent collaboration, where multiple conversational agents (CAs) embody distinct expert personas. For example, our study surfaced a key tension between maximizing exercise performance and preventing overexertion. To navigate such trade-offs, future systems could employ multiple agents (*e.g.*, a progressive planner focused on performance gains vs. a preventive planner emphasizing injury avoidance) that critique and refine plans from their respective perspectives. This setup would help users explore alternative viewpoints and make more informed decisions, addressing nuances that a single-agent model might overlook. This deliberative, agentic workflow has also been attempted in various domains to integrate complementary expertise and mirror real-world expert collaboration through discussion, negotiation, and consensus [12, 69]. By synthesizing diverse perspectives, we hypothesize that such systems involving multiple

agents with diverse viewpoints could generate more balanced plans and enhance informed decision-making.

7.3 Generalizability to Other Planning Domains

One key aspect of our system was the integration of implementation intentions, where the users are provided with `IF-THEN` statements linked to their availabilities collected through chatting with the CA. From the study, we identified that the participants perceived such situation-based expressions as highly comprehensible and adaptable, compared to vague amount-based or rigid time-based instructions. Similarly, as such implementation intention strategies have been shown effective in a variety of behavior change tasks (*e.g.*, diet control [2, 4, 33, 66], smoking cessation [16, 56]), we posit that our approach is also adaptable to various other behavior change contexts. Particularly, since our system is composed of a set of easy-to-alter instructions in a natural language that define the constraints to be collected, we believe that the adaptation process for various other tasks can be significantly straightforward, requiring minimal changes to tailor these instructions to reflect the domain-specific constraints of each new context.

7.4 Towards a Long-term Interaction with PlanFitting

From our formative study, we identified that exercise planning is an iterative process that takes place in the long term as the user's exercise progresses. Motivated by this, our exploratory user study simulated such an iteration and revealed opportunities for the CA as a long-term exercise companion. To further support the interaction of users with the CA in the extended duration, future works need to longitudinally study how the system can be expanded to help elicit information from users over time, and how to leverage that information to inform future revisions of the plan.

As the system scales up, we believe that incorporating context-aware features would help the plans to be even more aligned with the user-provided constraints, assisting in generating more realistic and customized plans. For instance, integrating location-aware exercise recommendations could enable PlanFitting to take into account factors driven by real-time information, such as weather conditions, nearby exercise facilities, or nearby routes that allow users to perform exercise on the go (*e.g.*, a specific route for running while going back home). Such a level of contextualization would make the generated plans even more closely connected to the user's real-world situation and make the exercise plans more engaging. Similarly, other features that reflect an up-to-date health status of the user could be incorporated into future revisions of PlanFitting to create even richer and more personalized exercise planning.

7.5 Limitations and Future Work

While our findings demonstrate the promise of LLM-driven conversational agents in personalizing exercise plans, several limitations remain, suggesting important directions for future work. First, our recommendations are based on a curated dataset that, while expert-vetted, may not reflect the full range of exercise types or cultural preferences. Future work should explore expanding the exercise corpus with more diverse, representative data sources, and develop mechanisms to detect and mitigate potential biases in both the dataset and model outputs.

Second, despite the use of retrieval-augmented generation and rule-based prompting, the LLM occasionally generated plans lacking nuance, such as missing rest days or offering overly general suggestions. These limitations point to the need for integrating more domain-specific reasoning or constraint satisfaction approaches alongside LLMs—via hybrid models or fine-tuning with expert-reviewed exercise prescriptions.

Third, our study focused on short-term interactions; long-term adherence, motivation, and engagement with AI-generated plans remain open questions. Future research should investigate how systems like PlanFitting perform in real-world, longitudinal deployments with usage over weeks or months, and explore interventions to sustain user motivation (*e.g.*, adaptive check-ins, habit formation scaffolds, social accountability features).

Finally, ethical and privacy considerations need to be further explored. Users may over-trust or misinterpret AI recommendations, especially in sensitive domains like health. Future versions should incorporate transparency mechanisms (*e.g.*, rationale generation, uncertainty estimation), offer opt-in controls over data sharing, and support human-in-the-loop oversight, ensuring that the agent is positioned as a supportive assistant rather than an authority.

8 CONCLUSION

In this study, we propose PlanFitting, an LLM-driven conversational agent that helps users create personalized exercise plans through natural dialogue. Based on a user study ($N = 18$) and evaluation of the generated plans, we highlighted PlanFitting's potential to guide personalized, guideline-informed exercise planning. We also discuss design implications for improving LLM-driven conversational agents in personalized exercise planning.

ACKNOWLEDGMENTS

We would like to thank Elena Agapie for generously providing the exercise dataset used in PlanFitting's exercise retrieval. We also thank participants from our formative interviews and user study for their time and effort. This work was supported through a research internship at NAVER AI Lab of NAVER Cloud.

REFERENCES

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *Journal of Medical Internet Research* 23, 1 (2021), e17828. https://doi.org/10.2196/17828

[2] Anja Achtziger, Peter M Gollwitzer, and Paschal Sheeran. 2008. Implementation Intentions and Shielding Goal Striving From Unwanted Thoughts and Feelings. *Personality and Social Psychology Bulletin* 34, 3 (2008), 381–393. https://doi.org/10.1177/0146167207311201

[3] ACSM. 2023. *The American College of Sports Medicine.* Retrieved Feb 1, 2025 from https://www.acsm.org/

[4] Marieke A Adriaanse, Charlotte DW Vinkers, Denise TD De Ridder, Joop J Hox, and John BF De Wit. 2011. Do Implementation Intentions Help to Eat a Healthy Diet? A Systematic Review and Meta-analysis of the Empirical Evidence. *Appetite* 56, 1 (2011), 183–193. https://doi.org/10.1016/j.appet.2010.10.012

[5] 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. https://doi.org/10.1145/3173574.3173898

[6] Elena Agapie, Lucas Colusso, Sean A Munson, and Gary Hsieh. 2016. PlanSourcing: Generating Behavior Change Plans with Friends and Crowds. In *Proceedings*

of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. 119–133. https://doi.org/10.1145/2818048.2819943

[7] 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. https://doi.org/10.1007/s11036-007-0011-7

[8] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep Me Updated! Memory Management in Long-term Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 3769–3787. https://doi.org/10.18653/v1/2022.findings-emnlp.276

[9] Niv Ben-Shabat, Gal Sharvit, Ben Meimis, Daniel Ben Joya, Ariel Sloma, David Kiderman, Aviv Shabat, Avishai M Tsur, Abdulla Watad, and Howard Amital. 2022. Assessing Data Gathering Quality of Chatbot Based Symptom Checkers-a Clinical Vignettes Study. *International Journal of Medical Informatics* 168 (2022), 104897. https://doi.org/10.1016/j.ijmedinf.2022.104897

[10] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706QP063OA

[11] Carlos Ivan Mesa Castrillon, Paula R Beckenkamp, Manuela L Ferreira, Jose A Michell, Vania Alice de Aguiar Mendes, Georgina M Luscombe, Emmanuel Stamatakis, and Paulo Henrique Ferreira. 2020. Are People in the Bush Really Physically Active? A Systematic Review and Meta-analysis of Physical Activity and Sedentary Behaviour in Rural Australians Populations. *Journal of Global Health* 10, 1 (2020). https://doi.org/10.7189/jogh.10.010410

[12] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848 [cs.CL]

[13] Zheng Chen. 2023. PALR: Personalization Aware LLMs for Recommendation. *arXiv preprint arXiv:2305.07622* (2023). https://arxiv.org/abs/2305.07622

[14] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-reflection: How People Reflect on Personal Data Through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 173–182. https://doi.org/10.1145/3154862.3154881

[15] Anna-Lisa Cohen and Peter M Gollwitzer. 2008. The Cost of Remembering to Remember: Cognitive Load and Implementation Intentions Influence Ongoing Task Performance. http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-61225

[16] Mark Conner and Andrea R Higgins. 2010. Long-term Effects of Implementation Intentions on Prevention of Smoking Uptake among Adolescents: a Cluster Randomized Controlled Trial. *Health Psychology* 29, 5 (2010), 529. https://doi.org/10.1037/a0020317

[17] 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. https://doi.org/10.1145/1124772.1124840

[18] Sunny Consolvo, Predrag Klasnja, David W McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A Landay. 2008. Flowers or a Robot Army? Encouraging Awareness & Activity with Personal, Mobile Displays. In *Proceedings of the 10th International Conference on Ubiquitous Computing*. 54–63. https://doi.org/10.1145/1409635.1409644

[19] Emely De Vet, Anke Oenema, and Johannes Brug. 2011. More or better: Do the number and specificity of implementation intentions matter in increasing physical activity? *Psychology of Sport and Exercise* 12, 4 (2011), 471–477. https://doi.org/10.1016/j.psychsport.2011.02.008

[20] Grace T DeSimone. 2019. The Tortoise Factor — Get FITT. *ACSM's Health & Fitness Journal* 23, 2 (2019), 3–4.

[21] J Larry Durstine, Benjamin Gordon, Zhengzhen Wang, and Xijuan Luo. 2013. Chronic disease and the link to physical activity. *Journal of Sport and Health Science* 2, 1 (2013), 3–11. https://doi.org/10.1016/j.jshs.2012.07.009

[22] Wilmer Stalin Erazo, Germán Patricio Guerrero, Carlos Carrión Betancourt, and Iván Sánchez Salazar. 2020. Chatbot Implementation to Collect Data on Possible COVID-19 Cases and Release the Pressure on the Primary Health Care System. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 0302–0307. https://doi.org/10.1109/IEMCON51383.2020.9284846

[23] Fitbit, Inc. 2024. Fitbit. Retrieved Feb 1, 2025 from https://www.fitbit.com

[24] Fitbod Inc. 2024. Fitbod. Retrieved Feb 1, 2025 from https://fitbod.me/

[25] FitnessAI Inc. 2024. FitnessAI. Retrieved Feb 1, 2025 from https://www.fitnessai.com/

[26] Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging Large Language Models in Conversational Recommender Systems. *arXiv preprint arXiv:2305.07961* (2023). https://arxiv.org/abs/2305.07961

[27] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[28] Carol Ewing Garber, Bryan Blissmer, Michael R Deschenes, Barry A Franklin, Michael J Lamonte, I-Min Lee, David C Nieman, and David P Swain. 2011. Quantity and Quality of Exercise for Developing and Maintaining Cardiorespiratory, Musculoskeletal, and Neuromotor Fitness in Apparently Healthy Adults: Guidance for Prescribing Exercise. (2011). https://doi.org/10.1249/MSS.0b013e318213fefb

[29] Garmin Ltd. 2024. Garmin. Retrieved Feb 1, 2025 from https://garmin.com

[30] Peter M Gollwitzer. 1999. Implementation Intentions: Strong Effects of Simple Plans. *American Psychologist* 54, 7 (1999), 493. https://doi.org/10.1037/0003-066X.54.7.493

[31] 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. https://doi.org/10.1016/S0065-2601(06)38002-1

[32] Google Inc. 2023. Gemini. Retrieved Feb 1, 2025 from https://gemini.google.com/

[33] Lucy Gratton, Rachel Povey, and David Clark-Carter. 2007. Promoting children's fruit and vegetable consumption: Interventions using the Theory of Planned Behaviour as a framework. *British Journal of Health Psychology* 12, 4 (2007), 639–650. https://doi.org/10.1348/135910706X171504

[34] Martin S. Hagger and Aleksandra Luszczynska. 2013. Implementation Intention and Action Planning Interventions in Health Contexts: State of the Research and Proposals for the Way Forward. *Applied Psychology: Health and Well-Being* 6, 1 (oct 2013), 1–47. https://doi.org/10.1111/aphw.12017

[35] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232* (2023). https://arxiv.org/abs/2311.05232

[36] Apple Inc. 2024. *Watch - Apple*. Retrieved Feb 1, 2025 from https://www.apple.com/watch/

[37] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 895–906. https://doi.org/10.1145/3196709.3196735

[38] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 15696–15707. https://proceedings.mlr.press/v202/kandpal23a.html

[39] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300316

[40] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21. https://doi.org/10.1145/3491102.3517457

[41] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–28. https://doi.org/10.1145/3130930

[42] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@Hand: Fostering Visual Exploration of Personal Data on Smartphones Leveraging Speech and Touch Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 17 pages. https://doi.org/10.1145/3411764.3445421

[43] 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*. 3063–3072. https://doi.org/10.1145/1978942.1979396

[44] Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. 2020. HarborBot: A Chatbot for Social Needs Screening. In *AMIA Annual Symposium Proceedings*, Vol. 2019. 552. https://europepmc.org/articles/PMC7153089

[45] Rafal Kocielnik, Raina Langevin, James S. George, Shota Akenaga, Amelia Wang, Darwin P. Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T. Hsieh, Kabir Yadav, Herbert Duber, Gary Hsieh, and Andrea L. Hartzler. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–10. https://doi.org/10.1145/3469595.3469599

[46] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: a conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous*

Technologies 2, 2 (2018), 1–26. https://doi.org/10.1145/3214273

[47] LangChain, Inc. 2024. LangChain. Retrieved Feb 1, 2025 from https://python.langchain.com/

[48] Min Kyung Lee, Junsung Kim, Jodi Forlizzi, and Sara Kiesler. 2015. Personalization Revisited: A Reflective Approach Helps People Better Personalize Health Services and Motivates Them To Increase Physical Activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 743–754. https://doi.org/10.1145/2750858.2807552

[49] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376175

[50] 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. https://doi.org/10.1145/1753326.1753409

[51] 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*. Springer, 261–278. https://doi.org/10.1007/11853565_16

[52] Sonia Lippke, Jochen P Ziegelmann, and Ralf Schwarzer. 2004. Initiation and Maintenance of Physical Exercise: Stage-Specific Effects of a Planning Intervention. *Research in Sports Medicine* 12, 3 (2004), 221–240. https://doi.org/10.1080/15438620490497567

[53] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374* (2023). https://doi.org/abs/2308.05374

[54] Aleksandra Luszczynska, Ralf Schwarzer, Sonia Lippke, and Magda Mazurkiewicz. 2011. Self-efficacy as a Moderator of the Planning—behaviour Relationship in Interventions Designed to Promote Physical Activity. *Psychology and Health* 26, 2 (2011), 151–166. https://doi.org/10.1080/08870446.2011.531571

[55] Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting Knowledge: Exploring the Creative Mechanisms of Chat-Based Search Engines. *arXiv preprint arXiv:2402.19421* (2024). https://arxiv.org/abs/2402.19421

[56] Lorna McWilliams, Sarah Bellhouse, Janelle Yorke, Kelly Lloyd, and Christopher J Armitage. 2019. Beyond 'Planning': A Meta-analysis of Implementation Intentions to Support Smoking Cessation. *Health Psychology* 38, 12 (2019), 1059. https://doi.org/10.1037/hea0000768

[57] Deana I Melton, Jeffrey A Katula, and Karen M Mustian. 2008. The Current State of Personal Training: an Industry Perspective of Personal Trainers in a Small Southeast Community. *Journal of Strength and Conditioning Research* 22, 3 (2008), 883. https://doi.org/10.1519/JSC.0b013e3181660dab

[58] U.S. Department of Health and Human Services. 2018. *Physical Activity Guidelines for Americans* (second ed.).

[59] National Academy of Sports Medicine. 2023. *How Much Does a Personal Trainer Cost & Should You Hire One?* https://blog.nasm.org/how-much-does-a-personal-trainer-cost

[60] OpenAI. 2022. Introducing ChatGPT. Retrieved Feb 1, 2025 from https://openai.com/index/chatgpt/

[61] OpenAI. 2024. API Platform. Retrieved Feb 1, 2025 from https://openai.com/api/

[62] OpenAI. 2025. *Function Calling*. Retrieved Feb 1, 2025 from https://platform.openai.com/docs/guides/function-calling

[63] Hyanghee Park and Joonhwan Lee. 2021. Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3411764.3445133

[64] Chelsea A Pelletier, Anne Pousette, Kirsten Ward, Robin Keahey, Gloria Fox, Sandra Allison, Drona Rasali, and Guy Faulkner. 2020. Implementation of Physical Activity Interventions in Rural, Remote, and Northern Communities: A Scoping Review. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* 57 (2020). https://doi.org/10.1177/0046958020935662

[65] 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. https://doi.org/10.1001/jama.2018.14854

[66] Tabea Reuter, Jochen P Ziegelmann, Amelie U Wiedemann, and Sonia Lippke. 2008. Dietary Planning as a Mediator of the Intention–Behavior Relation: An Experimental-Causal-Chain Design. *Applied Psychology* 57 (2008), 194–207. https://doi.org/10.1111/j.1464-0597.2008.00364.x

[67] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 3784–3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

[68] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (July 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[69] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. https://arxiv.org/abs/2306.03314

[70] Jared M Tucker, Gregory J Welk, and Nicholas K Beyler. 2011. Physical Activity in U.S. Adults: Compliance with the Physical Activity Guidelines for Americans. *American Journal of Preventive Medicine* 40, 4 (2011), 454–461. https://doi.org/10.1016/j.amepre.2010.12.016

[71] Timothy J Valshtein, Gabriele Oettingen, and Peter M Gollwitzer. 2020. Using Mental Contrasting with Implementation Intentions to Reduce Bedtime Procrastination: Two Randomised Trials. *Psychology & Health* 35, 3 (2020), 275–301. https://doi.org/10.1080/08870446.2019.1652753

[72] Nicole A VanKim and Toben F Nelson. 2013. Vigorous Physical Activity, Mental Health, Perceived Stress, and Socializing among College Students. *American Journal of Health Promotion* 28, 1 (2013), 7–15. https://doi.org/10.4278/ajhp.111101-QUAN-395

[73] Viswanath Venkatesh and Hillol Bala. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences* 39, 2 (2008), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[74] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *arXiv preprint arXiv:2301.05843* (2023). https://arxiv.org/abs/2301.05843

[75] Kefan Xu, Xinghui Yan, and Mark W Newman. 2022. Understanding People's Experience for Physical Activity Planning and Exploring the Impact of Historical Records on Plan Creation and Execution. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3491102.3501997

A LLM INSTRUCTIONS

A.1 Base Prompt for Dialogue Analyzer

```
- Analyze the input dialogue and return an array of JSON objects each of which denotes an update for this summary object.
- The user may mention multiple entities, such as goals and obstacles, or corrections to previous entities.
- You are allowed to use the following set of methods for update:
{
   target: "goal" | "availability" | "obstacle" | "recommended_exercise" | "implementation_intention",
   method: "add" | "update" | "remove"
   params: { // for update
     id: string,
     update: {} // will be overwritten to the corresponding element.
   } | { // for addition
     entity: {} // a new entity without ID; ID will be assigned by the system. Only for implementation_intention, assign a random
ID in case you use the "parent_ids" property.
   } | { // for removal
     id: string
   }
}
—
If there is nothing to be updated, return [].
```

B STUDY DETAILS

Table 4: Demographic information of participants in the formative study.

PID	Age	Gender	Occupation
FC1	44	Female	Homemaker
FC2	26	Female	Full-time employee (software engineer)
FC3	30	Female	Freelancer
FC4	56	Male	Retired
FC5	29	Female	Full-time employee (software engineer)
FC6	46	Female	Homemaker
FC7	31	Male	Graduate student
FC8	32	Female	Full-time employee (product manager)

Table 5: Demographic information of participants in the main user study.

PID	Age	Gender	Occupation	General goal for exercise
P1	39	Female	Full-time employee	Diet
P2	54	Male	Retired	Improve fitness, maintain muscle mass
P3	19	Male	Undergraduate student	Increase muscle mass, improve fitness
P4	45	Female	Homemaker	Manage blood pressure
P5	44	Female	Homemaker	Posture correction
P6	20	Male	Undergraduate student	Diet, improve fitness
P7	46	Female	Homemaker	Improve fitness, increase muscle mass, diet
P8	25	Female	Part-time employee (customer service)	Improve fitness, diet
P9	32	Female	Homemaker	Improve fitness
P10	19	Female	Undergraduate student	Improve fitness, diet
P11	27	Female	Part-time employee (sales)	Diet
P12	30	Male	Part-time employee (customer service)	Diet
P13	38	Female	Homemaker	Diet
P14	22	Male	Undergraduate student	Increase muscle mass
P15	24	Female	Undergraduate student	Increase muscle mass
P16	20	Male	Undergraduate student	Diet
P17	48	Male	Full-time employee (public sector)	Health management
P18	46	Female	Part-time employee (tax)	Increase muscle mass