

GPTCoach: Towards LLM-Based Physical Activity Coaching

Matthew Jörke
joerke@stanford.edu
Stanford University
Stanford, CA, USA

Shardul Sapkota
sapkota@stanford.edu
Stanford University
Stanford, CA, USA

Lyndsea Warkenthien
lyndsea@alumni.stanford.edu
Stanford University
Stanford, CA, USA

Niklas Vainio
niklasv@stanford.edu
Stanford University
Stanford, CA, USA

Paul Schmiedmayer
schmiedmayer@stanford.edu
Stanford University
Stanford, CA, USA

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

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

Nice to meet you, Alex!

I'm here to assist you in creating and maintaining a physical activity plan that suits your lifestyle and interests. In this program, you will be the one to design your own physical activity plan while I provide you with evidence-based support and feedback. The goal is to create a plan that reflects your preferences, resources, and goals.

This way, we can ensure that your plan is personalized, enjoyable, and sustainable for you. Do you have any questions or concerns about designing your physical activity plan or about the program in general?

I just became a new parent and am struggling to find time to exercise. I'm just not sure where to start

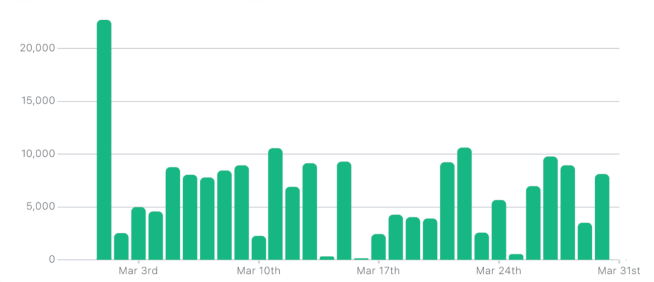
Alex, becoming a new parent is a major life stage and it's completely natural to feel overwhelmed. I can see why finding time for exercise could be a challenge. Remember, the key is to design a plan that fits into your current lifestyle and available resources. Don't worry, we are in this together and will find a way that works best for you. It's okay to start small, consistency is what matters the most.

...

Step Count

Day Week Month

March 2024



Thank you for sharing this information, Alex. Throughout March, you averaged around 6,738 steps per day, which is a good base to start from. This suggests that even amidst your busy schedule as a new parent, you've been engaging in some activities that involve movement.

Figure 1: GPTCoach is a health coaching chatbot designed to develop a physical activity plan that is tailored to the needs, abilities, and goals of a client. GPTCoach implements the onboarding conversation from Active Choices [66], an evidence-based health coaching program, uses counseling strategies from motivational interviewing, and can query and visualize a user's health data from a wearable device through tool use. On the left, we show an excerpt from an example conversation with GPTCoach that is representative of the conversation participants had in our lab study. On the right, we show an interactive visualization displayed by GPTCoach at a later point in the conversation.

ABSTRACT

Mobile health applications show promise for scalable physical activity promotion but are often insufficiently personalized. In contrast, health coaching offers highly personalized support but can be prohibitively expensive and inaccessible. This study draws inspiration from health coaching to explore how large language models (LLMs) might address personalization challenges in mobile health. We conduct formative interviews with 12 health professionals and 10 potential coaching recipients to develop design principles for an

LLM-based health coach. We then built GPTCoach, a chatbot that implements the onboarding conversation from an evidence-based coaching program, uses conversational strategies from motivational interviewing, and incorporates wearable data to create personalized physical activity plans. In a lab study with 16 participants using three months of historical data, we find promising evidence that GPTCoach gathers rich qualitative information to offer personalized support, with users feeling comfortable sharing concerns. We conclude with implications for future research on LLM-based physical activity support.

This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in:

CHI '25, April 26-May 1, 2025, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04...\$15.00

<https://doi.org/10.1145/3706598.3713819>

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI; Interactive systems and tools**; Natural language interfaces; • **Computing methodologies** → Natural language processing.

KEYWORDS

Physical activity, health coaching, large language models (LLMs), personal informatics, conversational agents

ACM Reference Format:

Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, New York, NY, USA, 45 pages. <https://doi.org/10.1145/3706598.3713819>

1 INTRODUCTION

Regular physical activity carries significant benefits to health [75, 103], yet over a quarter of the population worldwide [103] and nearly half of the population in the US [45] fail to meet standard recommendations for physical activity. The gold standard for encouraging health behavior change is one-on-one interaction with a human health professional [18]. Health coaching is an effective and common form of such personalized support, but it is expensive, not widely accessible, and does not scale [18, 91]. Researchers have turned to mobile health technologies such as smartphone applications or wearable fitness trackers as a low-cost, scalable approach for promoting active lifestyles [52]. One promise of mobile health applications is to use data from activity trackers and other sensors to personalize health interventions and support. For example, nearly all modern smartphones include a pedometer, which could be leveraged to tailor interventions that encourage walking. However, in comparison to support provided by health coaches, a key limitation of existing mobile health interventions is that they are insufficiently personalized to the *qualitative* aspects of a person's context, such as goals, values, preferences, past experiences, life circumstances, time constraints, physical abilities, or access to resources [96, 114]. These qualitative factors are generally the focus of coaching conversations [100, 113, 114, 132] and are crucial for effective support, but are challenging to standardize and quantify.

Recently, large language models (LLMs) have experienced rapid improvements in performance [20, 105], presenting new opportunities to address personalization challenges in mobile health. Recent work has explored applications of LLMs in health and medicine, including health inference tasks [42, 63, 78], conversational agents for diagnosis [1, 124], and mental health counseling [26, 117]. Inspired by these recent advances, we set out to explore the potential for LLMs to provide more personalized physical activity support than prior mobile health systems by drawing inspiration from human health coaching. Specifically, LLMs offer the potential to integrate various sources of context, including qualitative context captured in natural language (e.g., goals, life circumstances, or past experiences) and quantitative context from self-tracking data. Moreover, advances in conversational flexibility might allow a model to dynamically seek out information and adapt the structure and style of interaction in response to user input, much like a health coach.

Despite this promise, several technical and design challenges remain. Due to the probabilistic nature of LLM outputs and the fickleness of prompting [134], an LLM-based health coach requires a fundamentally different system design than existing rule-based chatbots. Moreover, off-the-shelf models do not natively support reasoning over raw sensor data as input [88] and are instruction-tuned to answer questions, not engage in open-ended coaching conversations. A burgeoning interest in LLMs for behavioral health within both academic research and commercial products has focused on health question-answering or insight-extraction tasks [37, 42, 44, 51, 78, 87, 101, 125]. However, these concurrent efforts offer limited guidance on designing LLM agents that can seek out and incorporate nuanced qualitative information about a user's personal and environmental context. Further, it remains unclear how to leverage these novel capabilities in ways that are endorsed by human health coaches and complements their coaching practice.

To navigate these challenges, we followed a human-centered process to design and evaluate GPTCoach, a GPT-4 based coaching chatbot for promoting physical activity. First, we conducted formative interviews with 22 participants, including 12 health professionals (health coaches, health educators, personal trainers, fitness instructors, and physical therapists) and 10 potential recipients of health coaching, ranging from highly sedentary individuals to athletes. Through a qualitative analysis of our formative interviews, we extracted three design principles that guided the design of GPTCoach: (1) follow a facilitative, non-prescriptive approach; (2) tailor information and advice using diverse source of context; and (3) adopt a supportive, non-judgmental tone. GPTCoach implements the onboarding conversation from the Stanford Active Choices program [66], an evidence-based, clinically-validated health coaching program [22, 65, 67, 130, 131] developed by experts in behavioral medicine. The onboarding conversation is one of the most important components of the Active Choices program and is crucial for effective support, setting the tone for the coaching relationship and requiring key skills such as building trust, maintaining a facilitative tone, and tailoring advice to a client's unique circumstances. GPTCoach also uses strategies from motivational interviewing (MI), an established, evidence-based counseling approach [89, 93], to engage in conversations that center the client's agency and motivation. Moreover, GPTCoach can query and visualize health data from a wearable through tool use [115]. At the end of the conversation, GPTCoach generates a personalized physical activity plan.

In a single-session lab study with 16 participants using three months of their historical self-tracking data, we evaluate GPTCoach's adherence to our design principles and participants' experiences interacting with GPTCoach. We hired trained human coders to evaluate GPTCoach's consistency with MI using the Motivational Interviewing Treatment Integrity (MITI) 4 code [94] and additionally perform a counterfactual comparison to GPT-4. Our single-session study design allowed us to evaluate GPTCoach's ability to conduct a high-quality onboarding conversation under full researcher supervision, mitigating potential ethical risks associated with unpredictable LLM outputs while providing valuable insights to inform future work on multi-session coaching.

Through our formative interviews, system design, and lab study, we make three key contributions to the literature. First, we contribute a set of design principles for LLM-based physical activity coaching, emphasizing a non-prescriptive, non-judgmental approach that tailors advice to each client's unique circumstances. Second, we contribute the system design of GPTCoach, including a novel prompt chaining [133] strategy to ensure appropriate adherence to the coaching program, use of MI strategies, and use of personal health data. Third, we offer insights and implications from our lab study evaluation of GPTCoach.

Our findings highlight the importance of considering the rich qualitative factors beyond sensor data that shape motivation, readiness for change, and the ability to take action. While concurrent work on LLMs for behavioral health has focused on extracting insights from wearable sensor data [37, 44, 88], our research demonstrates LLMs' potential to structure conversations in ways that centers the client's agency and fosters self-empowerment, which are critical aspects of successful health coaching programs [100, 132]. Survey and interview responses from our lab study indicate that participants perceived GPTCoach's guidance to be highly personalized and actionable, as well as that they felt supported by and comfortable sharing concerns with GPTCoach. Our MI coding results demonstrate that GPTCoach used MI-consistent or neutral behavior codes 93% of the time and outperforms vanilla GPT-4 in MI-consistency. However, GPTCoach's ability to use sensor data was more variable, sometimes demonstrating the capacity to use data in ways that served conversations about change and other times failing to proactively incorporate data into its advice. We conclude with a discussion of the implications for multi-session coaching, future LLM-based mobile health applications, and the risks of LLM-based health coaching.

2 RELATED WORK

In this section, we summarize relevant literature on health coaching, conversational agents for health behavior change, as well as personal informatics and self-reflection on personal data.

2.1 Health Coaching with Humans & Conversational Agents

Health coaching is a popular and effective intervention for motivating health behavior change [100, 132]. During health coaching, a coach establishes a partnership with their client and assists them in choosing healthy lifestyle behaviors by setting appropriate goals, maintaining accountability, and providing education and feedback [99]. Most health coaching programs advocate for a client-centric and non-prescriptive approach, including collaborative goal-setting, learning through self-discovery, fostering self-efficacy, and respecting client autonomy [132]. Many such programs draw from motivational interviewing [89], an evidence-based counseling framework that provides guidance on how coaches and counselors can facilitate conversations to elicit motivation for behavior change.

Studies on the role of wearable and self-tracking data in health coaching have found that data can provide coaches with more objective reports on the client's behaviors, which can aid in providing personalized care [31, 113, 114]. However, in a study examining

how clients and coaches interpret data during coaching sessions, Rutjes et al. [113] found that data were not "plug and play" and needed to be contextualized through collaborative reflection to inform behavior change. Moreover, coaches may lack the time or necessary expertise to interpret clients' data [31].

While highly effective, in-person health coaching is expensive, not always accessible, and does not scale [18, 91]. In contrast, *automated health coaching* [91] (also known as e-coaching [61]) aims to simulate the health coaching experience using conversational agents. Early work by Bickmore et al. on health dialogue systems [16, 18] has argued that automated systems can overcome the time and resource barriers affecting human counseling while effectively delivering evidence-based interventions at scale and low cost. In a meta-analysis of chatbots on lifestyle behaviors, Singh et al. [121] found a significant, small-to-moderate effect of chatbots on motivating physical activity. Many coaching chatbots draw from or mirror principles and strategies from human health coaching, such as motivational interviewing [5, 79, 98], and some also make use of external data sources [64, 69, 91]. Within HCI, several chatbot-based systems have been proposed for motivating physical activity [17, 33, 69, 92]. For example, Kocielnik et al. [69] developed an intent-based conversational system to explore mini-dialogues for triggering reflection on physical activity data. Mitchell et al. [91] compared a scripted Wizard-of-Oz health coaching chatbot to a human counseling condition for patients with type-2 diabetes, finding that chatbots led to human-like experiences despite their rule-based nature. Human coaches were more skilled at expressing empathy and tailoring support, while the chatbot was more persistent and consistent.

A limitation of all previously mentioned coaching chatbot systems is that they employ template-, rule-, or retrieval-based dialogue systems, which means that the system's response is always chosen from a pre-specified list of outputs. This offers a higher degree of researcher control but falls short of the conversational flexibility and degree of personalization afforded by human coaches. For example, in a review of conversational agents for physical activity promotion, Luo et al. [79] report that the most common challenges were related to the agent's capabilities, such as only allowing multi-choice response options. In contrast, LLM-based chatbots operate fundamentally differently by generating a unique response as a function of the conversation history. LLMs' responses mirror patterns learned in their internet-scale training data, which likely includes health blogs, medical guidelines, and coaching-related content, but may not always be reliable or up-to-date. LLM-based chatbots have been explored in other domains of health and medicine, including educational support [10], diagnosis [124], and mental health [24, 26, 55, 81, 119]. Emerging LLM-based approaches for physical activity coaching have explored prompting and re-ranking strategies to improve response quality [51] or using LLMs for health question answering [101, 125]. Recent work has also found that finetuned LLMs are capable of few-shot health tasks operating on raw self-tracking data [42, 78] and concurrent work has explored using LLMs for extracting insights from personal health data in the domains of physical activity and sleep [37, 44, 87]. Lastly, several commercially available LLM-based AI health coaches exist (e.g., WHOOP [129] or ONVY [102]), as well as the WHO's S.A.R.A.H [104]. Unlike our approach, all of the aforementioned

LLM-based systems focus on health question answering or insight extraction tasks and are thus not designed to seek out and incorporate nuanced qualitative information about a user's personal and environmental context. Informed by our expert interviews and the health coaching literature, we also place a greater focus on structuring open-ended conversations about change in ways that centers the client's agency and fosters self-empowerment.

2.2 Personal Informatics & Reflection on Personal Data

The field of personal informatics designs systems to “help people collect and reflect on personal information” [74]. Motivating physical activity is the most common application domain for personal informatics [43], with a long history of systems for motivating physical activity behavior change in HCI [7, 36, 76, 95]. Through reflection on personal data, personal informatics tools aim to generate new insights about behaviors and habits that can inform lasting behavior change [15, 27]. This has been referred to as the *self-improvement hypothesis* [62]. While the personal informatics literature has explored many design patterns for supporting reflection [15], many systems (and nearly all commercially available fitness trackers) make use of statistics and visualizations to support reflection [9, 28, 56, 86, 123].

Personal informatics tools are effective at fostering reflection and motivating behavior change [60], but many practical challenges remain. First, data interpretation is challenging for non-experts [48, 111]. In a diary study with non-expert users, Rapp and Cena [111] found that visualizations were “neither meaningful nor tailored to the user needs” and argued for greater personalization and contextualization of data. Moreover, highly quantitative forms of feedback may be potentially harmful [96], negatively impacting motivation [35] and mindsets [39]. In a review of self-tracking, Kersten-van Dijk et al. [62] found that insights gained through reflection on personal data are frequently not actionable towards behavior change. Many systems tacitly assume that reflection will occur naturally once data is processed and visualized [13], while most theories of reflection highlight that reflection needs to be explicitly supported [14, 122]. Our work aims to better support reflection on personal data by using LLMs to contextualize a user's health data to their unique qualitative context. Moreover, GPTCoach is instructed to ask reflective questions and assist the user in translating insights into actionable plans for change.

3 FORMATIVE INTERVIEW STUDY

Prior work has studied the role of self-tracking devices and data in the health coaching process [113, 114], as well as user experiences interacting with rule-based conversational agents [91]. However, it was unclear the degree to which these findings would transfer to LLM-based health coaching systems given the stark differences in capabilities. Thus, we conducted formative semi-structured interviews with 22 participants, including 12 health experts and 10 potential recipients of health coaching.

The goal of our formative study was to better understand how LLMs can provide personalized physical activity support by drawing inspiration from human health coaching practices. While human and automated coaching share many similarities, they differ

in their interaction medium and their respective capabilities. For example, as noted by Mitchell et al. [91], automated coaches are always available and consistent, while human coaches can establish personal connection through embodied interaction and shared lived experience. Recognizing these differences, we solicited diverse perspectives on the affordances and limitations of LLMs for health coaching. This approach allows us to understand where LLM health coaches might (or might not) excel and which aspects of human coaching should (or should not) be replicated.

Specifically, we investigated the following research questions:

RQ1: What strategies do health experts use to help their clients overcome barriers to physical activity? What strategies do individuals use to help themselves overcome barriers to physical activity? *Which of these strategies could LLMs employ and how?*

RQ2: How do health experts make use of clients' self-tracking data to promote physical activity? How do individuals make use of their self-tracking data to promote physical activity? *How might LLMs make use of self-tracking data to promote physical activity?*

3.1 Participants

We recruited 22 participants from various sources, including university mailing lists, personal contacts, and former research participants. 12 participants were health experts with professions including health coach, health educator, personal trainer, fitness instructor, YMCA vice president, physical therapist, and behavioral scientist. Six of the health experts had previous experience as peer health counselors in a study that evaluated a rule-based chatbot [64]. The 10 remaining non-expert participants were recruited from the general population to gain the perspective of potential recipients of health coaching. This participant pool included an NCAA Division 1 collegiate athlete, former athletes, avid self-trackers, physically active and inactive students, sedentary office workers, new parents, a program manager working on AI data quality, as well as active and inactive older adults. Participant demographics are listed in Table 1.

3.2 Procedure

We conducted one-hour semi-structured interviews with participants, pre-approved by our university's institutional review board. Health experts were asked several questions about their relationships with clients, strategies for overcoming activity barriers, and the role of technology and data in their practice. Non-experts discussed their physical activity levels, barriers, motivation sources, goals, and use of data. In the last portion of our interview, participants were asked to “imagine that you have access to an artificial intelligence chatbot that can help you improve your physical activity,” and were asked to share feedback on several potential features. Further details are provided in Appendix A.

3.3 Analysis

We performed qualitative coding on our interview transcripts using thematic analysis [21]. Two authors coded two interviews collaboratively, then independently coded roughly 20% of the interviews. One author coded the remaining 80%, with periodic meetings to

Age	Mean: 40.90, Median: 38, SD: 14.31, Min: 23, Max: 72
Gender	Female: 12, Male: 10
Race/Ethnicity	White: 10, Hispanic or Latino: 7, East Asian: 2, Southeast Asian: 2, South Asian: 1, African-American or Black: 1, Black English: 1

Table 1: Summary of participant demographics in the formative study ($N = 22$).

discuss codes and emerging themes, resolving any differences in interpretation through conversation.

3.4 Results

Towards answering **RQ1**, we discuss the role of coaches as facilitators, educators, and supporters. Towards addressing **RQ2**, we then discuss the role of data as guiding, not driving, physical activity behavior change.

3.4.1 RQ1: The Role of Coaches as Facilitators, Educators, and Supporters. We identified three major themes by analyzing the diverse strategies that coaches and individuals use to foster motivation and overcome barriers to behavior change. These themes center around coaches' roles in their interactions and relationships with clients. For each role, we discuss the implications for LLM coaching, including the affordances and limitations of LLMs in filling the role, as well as potential opportunities to go beyond the kinds of support provided by human coaches.

Facilitators: Despite our health experts having different training, job titles, responsibilities, and client populations, they all described an approach that was facilitative, not prescriptive in nature. Experts used various language to describe this strategy, such as *"You're not in the driver's seat, you're more in the passenger seat, providing maybe direction, steering the conversation one way or the other"* (P6). Core to a facilitative strategy is that the client takes ownership of their behavior change journey. A facilitative approach decidedly does not involve unsolicited advice or problem-solving: *"No, definitely not giving them the solutions [...] We're not advice givers"* (P15).

Experts cite several reasons for using facilitation as a strategy for overcoming barriers. First, several claimed that prescriptiveness does not work, as P12 reflected *"Friends that are like, give me a diet, or give me a workout plan. They never follow it."* Experts also noted that advising without comprehensive understanding and making assumptions could undermine credibility and trust. Facilitation empowers clients and fosters self-efficacy so that clients can sustain healthy behaviors and learn to solve problems on their own. As P11 emphasized, their facilitation goal with their clients was *"building that capacity so they can, so they can have the resiliency and tools needed to empower themselves."* Moreover, many coaches simply do not have time to adopt a prescriptive role: *"I want you to build those habits, [...] I don't have the time or energy for it. And also, I don't think that long term, that's the best strategy for building intrinsic motivation."* (P3).

Implications for LLM Coaching: Several experts independently cited motivational interviewing [89] as a guiding framework and many experts mentioned conversational strategies such as open-ended questions, reflective statements, affirmations, reframing, or advising with permission. Surprisingly, health experts did not describe

AI's behavior with facilitative terms, such as P15 suggesting that an AI could help *"come up with some solutions."* LLMs' conversational flexibility might allow them to structure facilitative conversations and implement strategies from frameworks like motivational interviewing, although this behavior generally conflicts with LLMs' instruction-following and questioning-answering objectives. This aligns with prior work in mental health support, which has found that LLM therapists prioritize providing solutions over asking questions [26].

Educators: Experts refrain from giving unsolicited advice and spoke about applying their advanced knowledge to help clients overcome barriers. As P8 noted, *"my superpower is sifting through a lot of information and pulling together the patterns."* This does not conflict with a facilitative approach; rather, information, education, and solutions should all be carefully tailored to a client's unique needs and situation: *"What is motivating you right now? [...] And then trying to find the common threads with things that [...] I know about and can help"* (P3). Open-ended questions, reflective listening, asking for permission, and other strategies assist coaches in ensuring that they have gathered enough information to provide advice that is aligned with the client's needs, abilities, and resources: *"The doctor said you need to be more physically active. So what I need you to do is go to a pool. [...] How do I know that you have access to a pool?"* (P19). Other kinds of education are targeted at reframing deep-seated beliefs about what counts as activity and why it is beneficial.

The process of tailoring advice is not an objective acquisition of information, as most experts also described engaging in interpretation: *"It's a mix of trying to provide what they necessarily want with what they need"* (P2). For example, experts and non-experts describe lack of time as an exceedingly common barrier to physical activity. While individuals differ greatly in their responsibilities and time commitments, experts noted that time barriers are also due to the perception of a lack of time. Thus, coaches aiming to assist clients in overcoming a lack of time might suggest both time-management strategies and help clients reframe their perceived time barriers.

Implications for LLM Coaching: LLMs have the potential to perform well—perhaps better than human experts—at providing personalized physical activity information, given their broad, internet-scale knowledge and advanced question-answering capabilities. Several participants appreciated this possibility: *"I think that is one of the most promising parts about generative AI, in my opinion, is being able to get the answer to somebody who doesn't necessarily know how to ask or how to look for it."* (P4). However, participants were also quick to acknowledge the limitations of LLMs' knowledge. Many experts were concerned about the AI's sources of information and potential for hallucination. Others noted that AI might perform

well for generic physical activity advice but could fail for highly specific activities.

Supporters: All of our health experts acknowledged the crucial role of personal connection and support, sharing comments like *“Just making everybody feel welcome. That’s it. No matter who you are, where you’re from, what your financial, social background is”* (P5). Health behavior change is a profoundly personal and emotional process. When asked to reflect on their barriers to physical activity, several of our non-expert participants shared traumatic experiences and insecurities, and experts mentioned that many of their clients lack confidence or have anxieties. The strongest differentiator between our most active and inactive participants revolved around identity—whether being active was core to who they were. Highly active participants used language that affirmed this identity, *“It’s just in my DNA. It’s just what I love to do”* (P5), while many of the most challenging barriers were rooted in identity conflict: *“It’s pretty depressing some days, you know, it’s like I’m missing half of myself [...] I’m just a mom, and then I think back on those days when I did skate and compete pretty regularly”* (P10).

Health coaching is fundamentally relational—a partnership between a coach and client working towards a shared goal [100]. Experts emphasized the importance of building rapport and trust with their clients as a way to reduce anxiety and fear: *“once I meet with the clients and we kind of develop a little bit of rapport, it kind of takes a little bit of animosity from the weight room”* (P2). Many coaches highlight positive affirmations, such as *“the strategies that I always use for this population is a lot of encouragement, a lot of celebration”* (P1). These positive gestures can not only make clients more comfortable, but also encourage accountability and habit formation: *“I think accountability is really important, having somebody at your corner and feeling that support”* (P6). Each coach supports dozens, if not hundreds, of clients, and there are limits to coaches’ capacities: *“What I’ve had to learn is I can’t help everybody, and not everybody can afford to see me, and I feel terrible about that”* (P8). Many experts emphasized the importance of building community to encourage mutual support and reduce reliance on the coach.

Implications for LLM Coaching: Expert and non-expert participants had mixed opinions on whether an AI can or should establish such personal connections. Most people liked the ability of an AI to hold them accountable with personalized and encouraging reminders. Many also liked the ability for an AI to adapt its persona, e.g., *“depending on the day, I will need a different person. Sometimes I’m going to need that empathetic, but other times I’ll need the strict regimen”* (P13). However, many hesitated to discuss highly personal, emotional topics with an AI, such as P10 disclosing *“I guess that’s a little too personal. [...] Are we here for exercise, or are we here for, like, mental counseling?”* Some did not believe it was possible to form such a connection with an AI: *“I don’t think that my type of job, instructor wise, will ever be taken [...] Even though it will have all the information, it’s not personal.”* (P5), while others were readily open to the idea. Interestingly, a health coach who had previously worked on a rule-based chatbot mentioned that *“when you play video games, you immerse yourself in a world [...] these are older adults that have never been exposed to anything like this before, and I think they would develop a connection”* (P11). Moreover, this chatbot had advantages over human coaches: *“they had this avatar that they*

can go see once a week, was always there, like, didn’t run late, you know, didn’t judge them” (P11).

In line with prior literature, our findings suggest that people can and will form personal connections with a chatbot (even if they deny it [112]) and that this can be leveraged to create accountability and motivation toward physical activity behavior change with positive affirmations and encouragement. We expect that LLMs will offer marked improvements over rule-based chatbots. While LLMs can project empathy [40], it is clear that LLMs **should not be designed to replace human connection**, but rather to foster motivation and bolster existing relationships with coaches and communities.

3.4.2 RQ2: The Role of Data & Technology as Guiders, not Drivers. Many of our participants used wearable trackers and apps. In line with prior work in personal informatics [29, 30], participants cited many reasons for using data, including monitoring daily fluctuations, long-term trends, progress towards goals, or in-the-moment workout statistics. Participants appreciated wearable data as a more “objective” account of behavior, but were particularly frustrated when their behavior was not captured [36], like P5 mentioning *“If I forget to put my watch on, my whole day is ruined. [...] the day just doesn’t exist anymore.”* Most health experts acknowledged the utility of self-tracking data and characterized data as most helpful in maintaining accountability and consistency for individuals already on the path to becoming active. By analogy, one health educator asked us, *“How would driving a car be a different experience if you had no gauges in front of you?”* (P18)

Despite this, several coaches did not analyze data with their clients except in high-performance athletic or clinical contexts. Many coaches do not have time to analyze data, *“I can’t scale that. I have, like, 20 clients”* (P3), and also acknowledged their biases, *“Despite me having sort of more understanding of fitness science or exercise science broadly, I’m still falling into the same traps”* (P4). Incorporating data into health coaching also presents additional challenges. Data showing a lack of progress can negatively impact motivation and data often lacks important context: *“I feel like a lot of people, especially in fitness, think very quantitatively for everything, which is good to a certain extent, but, like, doesn’t give you the whole picture”* (P1).

Implications for LLM Coaching: Both expert and non-expert participants were generally excited by the prospect of an AI coach analyzing data. Participants had a dominant view of AI as an information-synthesis machine, such as P5 stating *“With AI, it has so much information within it. All you would have to do is tell them what your goals are, and it will obviously give you an answer.”* Participants wanted an AI coach to use their data to help them set goals and maintain accountability: *“I’m not accountable to anyone other than myself, [...] but, if you had a chatbot going hey, you only got 7000 steps and now you’re sitting on your butt”* (P7). An AI coach could also help flexibly re-adjust goals over time to changes in motivation, life circumstances, or injuries. Finally, many participants expressed privacy concerns when integrating health data with a chatbot. Even when data is fully secured, experts noted that user perceptions of privacy and prior expectations play a critical role: *“it’s about what people think about what’s going to happen. This is about people’s perceptions. [...] Family and cultural dynamics that come into play also trust can also play a big role”* (P19).

LLMs' use of wearable data is an opportunity to better support health behavior change. Conversational LLMs could enable personalized data analysis in ways tailored to the client, such as reflective sessions that examine historical data, establishing baselines for goal-setting, providing real-time support during or after workouts, or augmenting missing/inaccurate data with additional context. To provide adequate data-driven support, an agent must integrate data with extensive qualitative information about a user's life and carefully present this information to foster motivation and empowerment.

4 GPTCOACH: DESIGN & IMPLEMENTATION

Guided by our formative interviews, we designed GPTCoach, a chatbot system for physical activity coaching. In this section, we discuss our design process and system architecture. Our code is available at: <https://github.com/stanfordhci/gptcoach-chi2025>.

4.1 Design Principles

Our formative interviews revealed the role of coaches as facilitators, educators, and supporters, as well as insights into how LLMs might fill each of these roles. From these roles, we extracted three design principles for GPTCoach:

- DP-1: Follow a facilitative, non-prescriptive approach.** Health experts all described a facilitative approach to coaching in which clients drive their own behavior change journey. The chatbot should similarly stay “in the passenger seat,” empowering clients to make a change rather than prescribing what to do.
- DP-2: Tailor information and advice using diverse sources of context.** Any information and advice provided by the chatbot should be tailored to a user's unique personal and environmental factors. The chatbot should integrate both qualitative and quantitative sources of context.
- DP-3: Adopt a supportive, non-judgmental tone.** Health behavior change is a deeply personal journey and many clients face anxieties and fears around exercise. The chatbot should adopt an uplifting, non-judgmental tone to help clients feel comfortable and supported.

4.2 The Active Choices Program

We designed GPTCoach to implement the onboarding conversation of the Stanford Active Choices program [66], an evidence-based, clinically validated counseling program for physical activity promotion [22, 65, 67, 130, 131]. Active Choices is grounded in behavior change theory, including the transtheoretical model [110] and social cognitive theory [12]. During onboarding, coaches first introduce themselves and outline program expectations. They then discuss the client's past experiences, barriers to physical activity, health or injury concerns, and motivations. The session concludes with collaborative goal-setting based on the FITT (Frequency, Intensity, Time, and Type) framework along with advice (with permission, and if appropriate) for helping the client achieve their goal. Although other health-related topics may arise (e.g., nutrition, weight loss, sleep, and mental health), Active Choices coaches are encouraged to steer the discussion back to physical activity. We adopt a

similar approach in GPTCoach, treating these topics as largely out of scope, but not entirely off-limits.

In the full Active Choice program, the onboarding conversation is followed by several shorter follow-up contacts that occur every few weeks over six months or more. In this work, we focus only on the onboarding conversation to evaluate GPTCoach's adherence to design principles in a single-session lab study. Given the open-ended and highly personal nature of health coaching, the risks of LLM coaching are greater than traditional mobile health applications. This study design allows for full researcher supervision and intervention if necessary, minimizing potential risks from unpredictable outputs. Although multi-session coaching is generally required for sustained behavior change, we note that the onboarding conversation is the longest and potentially most impactful session, setting the tone for the program and the relationship with the client. Successful onboarding conversations require key skills that are transferable to follow-up sessions, such as facilitating a supportive and non-judgmental conversation, seeking out important information about a client's background and motivation, and tailoring information and advice to a client's unique circumstances. Thus, a system capable of conducting high-quality onboarding conversations is likely to perform well in follow-up contacts. We explore how GPTCoach might be extended for follow-up sessions in Section 8.

4.3 Design Process

LLMs present several design challenges due to their unpredictable outputs and the fickle nature of prompting [134]. Thus, we designed GPTCoach as a *technology probe* [57] to assess the viability of our approach and generate new ideas for design.

We engaged in several design activities in creating GPTCoach. We received training manuals from the Active Choices team along with descriptive statistics from a previous study on a rule-based chatbot [64]. During initial experiments, we utilized our own wearable data to develop a prototype resembling the final architecture, but lacking an additional prompt chain for tool calls (Section 4.4.2). This prototype was tested with two health experts and four non-experts in a pilot study. Feedback from participants highlighted a need for the model to more proactively query for data. In response, we adjusted nearly all of our prompts and added the tool call prompt chain to explicitly check for opportunities to query data. Since our prompt chaining approach proved to be effective in our pilot studies, we did not explore finetuning or preference alignment [105] in this work, which would necessitate an additional dataset of expert annotations.

4.4 System Architecture

Our system's architecture (Figure 2) consists of four main components: (1) a *database* containing raw health data and conversation histories; (2) an *iOS application* that fetches three months of historical data using Apple's HealthKit API and uploads the data to our database; (3) a *backend server* (Python), which handles all LLM logic and tool call execution, and (4) a *frontend web interface* (TypeScript/React) that displays the chat interface and interactive data visualizations. Our system uses Google Cloud Firestore for our database and Google Cloud Run to host the backend and frontend.

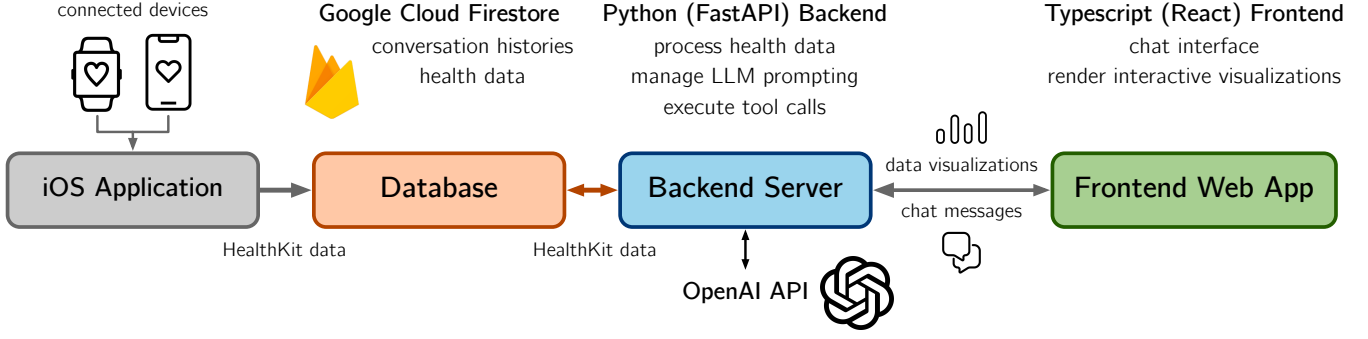


Figure 2: Overview of GPTCoach's System Architecture. HealthKit data from connected devices (e.g., iPhones and wearables) are synced to our Google Cloud Firestore Database using an iOS application. A Python backend server handles several features: fetching health data from the database, aggregating and featurizing health data, handling LLM prompt management logic, interfacing with the OpenAI API to query GPT4, executing tool calls, generating data visualizations for the frontend, as well as sending and receiving chat messages from the frontend. The Typescript/React frontend web app hosts the chat interface and renders interactive data visualizations.

GPTCoach builds on the open-source Spezi ecosystem [116] for creating iOS-based digital health apps and uses GPT-4 [4] via the OpenAI Chat Completions API for LLM interactions. In the following sections, we discuss how our architecture supports a data pipeline that enables our chatbot to call tools that fetch personal health data, as well as several prompt chains [133] to encourage adherence to the coaching program, adherence to motivational interviewing, and appropriate use of wearable data.

4.4.1 Data Pipeline. Our data pipeline consists of a set of data source inputs (e.g., heart rate or step count). The exact set of data sources depends on the user's device(s) and permissions granted to our application. All possible data sources are listed in Appendix B. We use the Spezi framework [116] (iOS/Swift) to read data from Apple HealthKit, encode it into the HL7 FHIR standard [50], and upload it to Firebase. Though we make use of Apple HealthKit to fetch health data, our system is not limited to Apple devices. Any wearable device that syncs with HealthKit (e.g., Oura and WHOOP) can be used as input to our system.

To provide our model access to information from wearable sensor data, we expose two tools to the LLM:

- **describe(data_source, date, granularity):** This function fetches all data within the given granularity ('day|week|month') from the reference date and returns a natural language description containing aggregated summary statistics and a data source description.
- **visualize(data_source, date, granularity):** This function returns the same output as describe, but additionally sends a message to the frontend to display an interactive data visualization to the user (see Figure 1B).

The summary statistics and visualizations reported by the function call depend on the type of data source. For instance, count types, such as step count, active energy burned, or exercise time, are summed and visualized with a bar chart. Rate types, such as heart rate, oxygen saturation, and respiratory rate, are averaged and visualized with a line chart. Our system is extensible in that new types and visualizations can be added flexibly for new data sources.

A tool call is initiated by GPT-4, which determines when and how to trigger a function based on the input. The Chat Completions API generates a JSON output based on the function signature, specifying the input arguments and their types. When the model initiates a tool call, the backend fetches the data from the database, generates the aggregated summary statistics, and returns a summarized text description to the model. For example, our data pipeline would return the following text description to the LLM after generating the visualization in Figure 1:

```
> describe(
  data_source_name="health.stepcount",
  start="2024-03-01",
  end="2024-03-31",
  granularity="day"
)

2024-03-01:00:00:00 to 2024-03-01:23:59:59:
22728 steps from Apple Watch (108 entries)
2024-03-02:00:00:00 to 2024-03-02:23:59:59:
2528 steps from Apple Watch (21 entries)
2024-03-03:00:00:00 to 2024-03-03:23:59:59:
4987 steps from Apple Watch (13 entries)
...
2024-03-29:00:00:00 to 2024-03-29:23:59:59:
3510.00 steps from Apple Watch (60 entries)
```

4.4.2 Prompt Chains. We encountered several limitations to vanilla prompting approaches¹. The model struggled to adhere to the onboarding session's structure, easily veered off-course, and had a strong tendency to give unsolicited advice. We provide further quantitative evidence of vanilla GPT-4's bias towards unsolicited advice in Section 6.5. To address these issues, we constructed three prompt chains [133] to elicit our desired behaviors: a *dialogue state* chain, a *motivational interviewing* chain, and a *tool call* chain. Upon receiving a user message, the conversation history sequentially

¹We define "vanilla prompting" as single-step output generation where a model f samples a response $y \sim f(\cdot|x)$ conditioned on a prompt x . In contrast, a prompt chain relies on several intermediate generations $x_t \sim f(\cdot|x_1, \dots, x_{t-1})$ to produce the final response $y \sim f(\cdot|x_1, \dots, x_n)$.

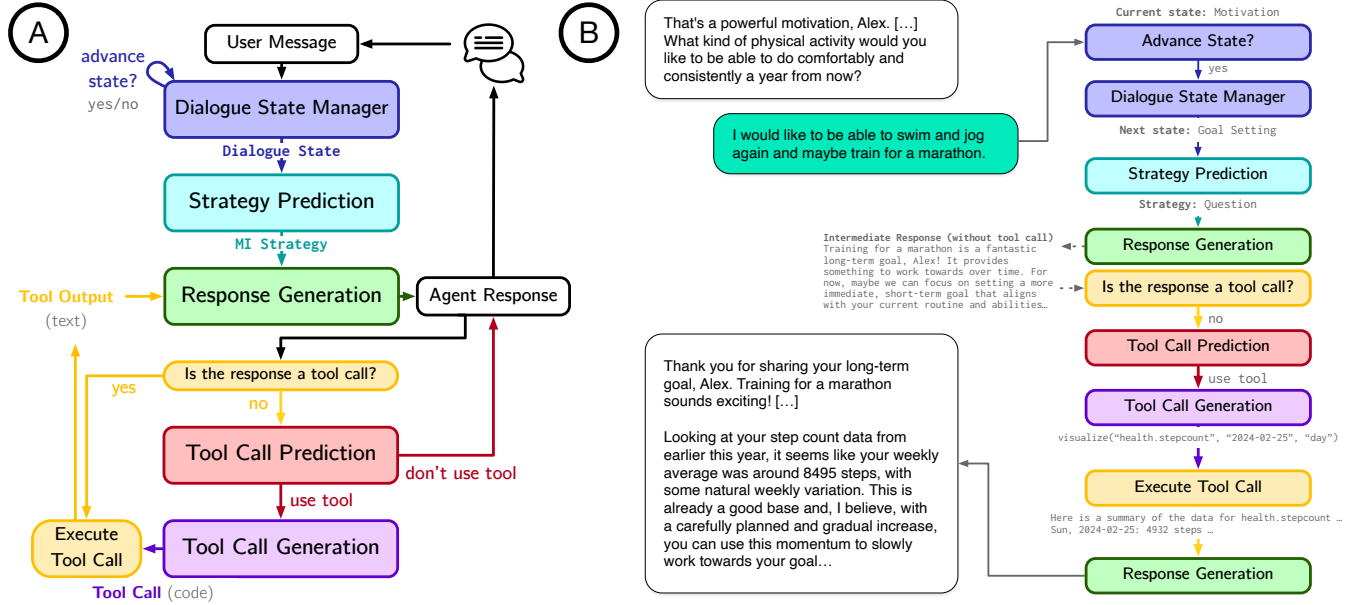


Figure 3: Overview and a walkthrough of GPTCoach’s Prompt Chains. (A) On the left, we show an overview of the prompt chains, in which the first chain manages the dialogue state, the second chain grounds the model’s response in MI strategies, and the third chain determines whether the response should be augmented with health data. (B) On the right, we show the outputs of each prompt chain for an example exchange.

passes through these chains (see Figure 3), each of which initiates a separate call to a GPT agent. We describe each prompt chain below and provide the full structure of all prompts in Appendix B.2.

Dialogue State Chain. The onboarding session consists of a series of topics that the coach discusses with the client. However, the agent needs to maintain flexibility throughout the conversation. For example, if a client asks an off-topic question, the coach should gently redirect the conversation back on topic. If a client gives an incomplete answer, the agent should politely follow up. Similarly, if a client mentions an injury early in the session, the agent should either reference this concern when asking about health or injury concerns or skip this question.

We organized the session into a linear sequence of dialogue states, each associated with a prompt. We created these prompts based on our formative interviews with health experts, as well as the Active Choices manual. Each prompt consists of a clear high-level task (e.g., “Your current task is to help your client set a physical activity goal.”), along with state-specific subtasks (e.g., “First, help them set a short term goal, if they have not already identified one themselves.”) and advice (e.g., “Connecting their short term goal to larger motivations can help them stay motivated.”). When the user sends a message, an external LLM agent classifies whether “the agent has successfully completed the following task,” advancing to the next dialogue state if the task is complete. All dialogue state prompts are provided in Appendix E, Figure 17.

Motivational Interviewing Chain. While our dialogue state chain manages what the model should talk about, it offers little guidance on how it should say it. We selected 11 motivational interviewing strategies from the Motivational Interviewing Skills Code

(MISC) [93], filtering out codes corresponding to undesired counselor behavior, merging some codes to reduce overlap, and adapting the examples to the domain of physical activity coaching. The 11 codes are listed in Table 2. We use another prompt chain, inspired by [55, 106, 118], to ground the model’s behavior in motivational interviewing strategies: one agent selects an MI strategy conditioned on the current dialogue state and history, another then generates a response following this strategy.

Tool Use Chain. If the response generation step does not call a tool, we prompt an additional tool call prediction agent to “determine whether the agent’s response should be augmented with the user’s health data.” If yes, we force the model to generate an output that calls the visualize function.

5 EVALUATION STUDY

We evaluate GPTCoach as a technology probe in a lab study with 16 participants. Our study was approved by our university’s institutional review board. As Klasnja et al. [68] have argued, assessing behavior change in a traditional sense (e.g., via a longitudinal RCT) is often an inappropriate metric for early-stage technologies in HCI research. Instead, as is common in HCI research on systems targeting health behavior change, we focus our evaluation on participants’ experiences and GPTCoach’s adherence to our design principles.

5.1 Participants

We recruited 16 participants from the general population using a variety of sources, including university mailing lists, social media

Strategy	Description	Example
ADVISE WITH PERMISSION	Give advice, make a suggestion, or offer a solution or possible action, after gaining permission.	“Would it be all right if I suggested something?”
AFFIRM	Say something positive or complimentary to the client.	“You’re a very resourceful person.”
FACILITATE	Simple utterances that function as “keep going” acknowledgments.	“Hmm. Tell me more.”
FILLER	Responses not categorizable elsewhere, such as pleasantries.	“Good morning, John.”
GIVING INFORMATION	Give information to the client, explain something, educate or provide feedback or disclose personal information.	“Your heart rate was higher during today’s workout.”
QUESTION	Ask a question in order to gather information, understand, or elicit the client’s story.	“How do you feel about that?”
RAISE CONCERN	Point out a possible problem with a client’s goal, plan, or intention.	“I’m worried about your plan to decrease workout days.”
REFLECT	A reflective listening statement in response to a client statement.	“You’re looking for a more relaxed environment.”
REFRAME	Suggest a different meaning for an experience expressed by the client, placing it in a new light.	Client: “My husband is always nagging me about going to the gym.” Counselor: “It sounds like he’s concerned about your health.”
SUPPORT	Generally sympathetic, compassionate, or understanding comments.	“That must have been difficult.”
STRUCTURE	Give information about what’s going to happen directly to the client throughout the course of treatment or within a study format.	“What we normally do is start by asking about your physical activity.”

Table 2: Motivational interviewing strategies used by GPTCoach in the motivational interviewing prompt chain. The strategies were adapted from the Motivational Interviewing Skills Code (MISC) [93]. The model first picks which of the 11 strategies above to use and then generates a response conditioned on the chosen strategy.

Age	Mean: 38.2, Median: 32.5, SD: 14.8, Min: 21, Max: 71
Gender	Female: 10, Male: 6
Race/Ethnicity	White: 10, Hispanic or Latino: 2, Southeast Asian: 2, African-American or Black: 1, East Asian: 1, Middle Eastern: 1, South Asian: 1
Education	Associate: 2, Bachelor’s: 6, Master’s: 7, Doctorate: 1
Stage of Change	Precontemplation: 1, Contemplation: 8, Action: 4, Maintenance: 3
Level of Activity (IPAQ)	Low: 5, Moderate: 7, High: 4
AI Knowledge	Novice: 1, Basic: 10, Intermediate: 3, Advanced: 2

Table 3: Summary of participant demographics in the technology probe evaluation study ($N = 16$).

advertisements, former research participants, and personal connections. Participant demographics are listed in Table 3. Participants were required to own an iPhone (as our system relies on Apple HealthKit) and nine also owned an Apple Watch. All participants were required to be fluent in English. Nearly all participants (15/16) expressed interest in increasing their physical activity in the near future (one participant responded “I don’t know”), aligning with our study’s focus on behavior change. We discuss the limitations of our recruitment procedure and potential biases in Section 7.

Participants were selected using a screening survey for a balanced sample across several dimensions, including age (21–71, Mean: 38.2, SD: 14.8), gender (10 female, 6 male), exercise stage of change [83] from the transtheoretical model [110] (1 precontemplation, 8 contemplation, 4 action, 3 maintenance), and levels of physical activity assessed via the International Physical Activity Questionnaire (IPAQ) [38] (5 low, 7 moderate, 4 high). The majority of participants (13/16) were employed for wages and varied

in their level of education, including associate degrees (2), bachelor's degrees (6), master's degrees (7), and one doctorate. Two participants self-identified as having a disability. Most participants (14/16) reported having previously interacted with an AI chatbot. Their knowledge of AI varied, with one participant identifying as a novice, the majority (10/16) having basic knowledge, three participants having intermediate knowledge, and two reporting advanced knowledge.

5.2 Procedure

Participants interacted with GPTCoach in a one-hour study session, including both in-person (5/16) and remote sessions conducted on Zoom (11/16). We did not encounter notable differences between the two formats and report usability, advice quality, and demographics by format in Appendix C.2. Prior to the session, participants uploaded three months of historical data from HealthKit using our iOS application (Appendix B.1). All participants were informed that a chatbot would have access to their health data in our study's consent form and were reminded again at the beginning of the session.

At the beginning of the session, participants were informed that they would be interacting with a chatbot acting as a health coach and that this conversation would mirror an onboarding conversation in an established health coaching program. They were told to interact with the system as they usually would while thinking aloud. There was no training and no further instructions were provided. If questions arose about using the chat interface, they were answered; otherwise, participants were reminded to interact as they normally would. The session concluded with open-ended questions from the researcher about their overall experience and a post-study survey.

5.3 Analysis

Our analysis uses several methods, including survey measures, qualitative coding using thematic analysis, computational analysis of model states and transcripts, and motivational interviewing coding by human experts.

5.3.1 Survey Measures. We asked participants several 5-point Likert scale questions about their experience interacting with GPTCoach and the quality of its advice. In addition, we measured usability using a subset of the Subjective Assessment of Speech Interfaces [53] (the same subset as Mitchell et al. [91] used with two additional questions from the habitability and speed factors). All questions are provided in Appendix C.1.

5.3.2 Qualitative Coding. To better understand participants' experiences with GPTCoach and GPTCoach's adherence to our design principles, we qualitatively coded interviews (including conversation transcripts, think-alouds, and post-study interviews) using thematic analysis [21], using a similar process that was used in the formative studies (Section 3.3).

5.3.3 Dialogue Analysis. We report the overall frequency and temporal progression of GPTCoach's internal dialogue states and MI strategies (as used in the prompt chains), along with an analysis of tool calls. We also report differences in message length.

5.3.4 Motivational Interviewing Coding. While each response is generated conditioned on a single MI strategy, the agent frequently makes use of several strategies within a single response. For example, this response was generated conditioned on the QUESTION strategy, but the agent begins the response with GIVING INFORMATION: "*Starting and maintaining motivation for physical activity can indeed be challenging, and you're not alone in feeling this way. Now, have you ever had any health problems or injuries that have interfered with your ability to exercise?*" Moreover, GPTCoach may or may not implement a given MI strategy in a way that is consistent with MI principles.

To better assess GPTCoach's consistency with MI, we performed motivational interviewing coding on all participant conversations with GPTCoach. As a first step, we performed LLM-based coding following Chiu et al.'s [26] prompt-based method for evaluating LLM psychotherapists, using the MISC coding scheme [93] (full details provided in Appendix D.2). To more rigorously evaluate our system, we then partnered with an MI agency and hired six trained, experienced coders to perform behavior coding on all conversation transcripts according to the Motivational Interviewing Treatment Integrity (MITI) Code 4 [94]. While we used MI strategies from MISC [93] in designing our MI chain and initial LLM-based coding, we transitioned to MITI for human coding based on expert advice. MITI is a behavioral coding system derived from MISC that was specifically designed to measure MI fidelity with a simpler and more reliable codeset. MITI is commonly used in prior work [90, 107–109, 117, 127, 128] and is used by the agency to evaluate human therapists. Each transcript was independently coded three times, with inter-rater reliability indicated by ICC scores ranging from 0.44 to 0.98 and an average (SD) ICC of 0.79 (0.17), reflecting a high overall level of agreement [32]. Full details on the annotation process and inter-rater reliability are provided in Appendix D.1.

MITI defines 10 utterance-level *behavior codes* that capture various counselor behaviors, which are highly analogous, but not exactly identical to the MISC codes listed in Table 2. Following Miller and Rollnick [89, p. 289], we consider the codes PERSUADE and CONFRONT to be *MI-inconsistent*. We consider GIVING INFORMATION to be *neutral* and the remaining 7 codes to be *MI-consistent*. We report the frequency of each behavior code as well as the overall frequency of MI-consistent, neutral, and inconsistent codes. Moreover, we report a summary of qualitative feedback provided by the trained coders. We refer to the MISC strategies (Table 2) used in the GPTCOACH's motivational interviewing chain as *internal* MI strategies and the MITI behavior codes annotated by human experts (Figure 12; Table 12) as *external* MI strategies.

5.3.5 Counterfactual Analysis. We perform a counterfactual analysis to compare our model's behavior to GPT-4 with vanilla prompting. We condition both models on the first five turns of each participant's conversation, covering program introduction, participants sharing their name and age, and ending with the agent asking for questions or concerns. For each of the 16 participants, we simulate 10 responses representing different barriers to physical activity based on coaching materials from collaborators in behavioral medicine. For each of the 160 simulated histories, we generate outputs using GPTCoach and using GPT-4 with only the system prompt and all prompt chains removed (which we refer to as "vanilla GPT").

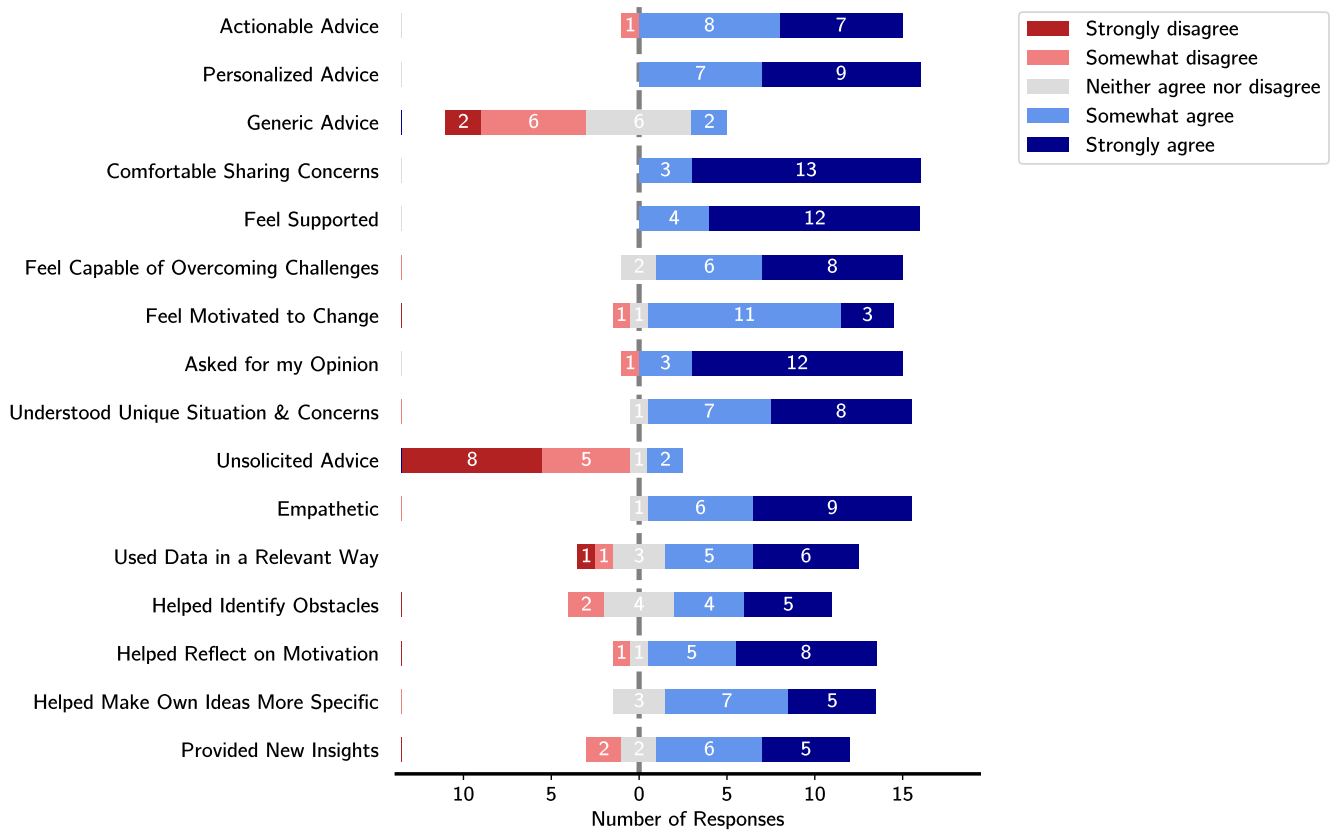


Figure 4: Participant Responses to Survey Items on User Experience & Quality of Advice. Participants had an overwhelming positive, comfortable, and supportive experience interacting with GPTCoach. The advice they received was personalized, actionable, and not unsolicited. Full questions are provided in Appendix C.1.1.

Subsequently, we evaluate all responses for MI consistency by having a single coder perform MITI behavior coding on each of the 320 total messages. The coder was the most experienced coder in our pool and was blind to condition. Full details on the counterfactual analysis along with illustrative examples are provided in Appendix D.1.3.

6 RESULTS

In this section, we report on the result of our survey measures, qualitative analysis, computational analysis of conversation transcripts, and MI coding evaluation.

6.1 Survey Measures

6.1.1 User Experience & Quality of Advice. Figure 4 illustrates that participants had an overwhelmingly positive experience with GPTCoach. On average, they felt supported (4.8/5; average response from 1: *Strongly disagree* to 5: *Strongly agree*), comfortable sharing concerns (4.8/5), capable of overcoming challenge (4.4/5), and motivated to change (4.0/5). The advice provided was rated as personalized (4.6/5), actionable (4.3/5), not unsolicited (1.8/5), and (to a lesser extent) not generic (2.5/5). The chatbot was empathetic (4.5/5), and helped them reflect on motivation (4.3/5). To a lesser extent,

the chatbot used data in a relevant way (3.9/5), helped identify obstacles (3.8/5), helped make their own ideas more specific (3.9/5), and provided new insights (3.7/5).

6.1.2 Usability. Among the subset of 12 questions selected from SASSI [53] (see Appendix C.1.2), our usability evaluations yielded an aggregated score of 49.4/60 (82.4%). The questions are grouped into six factors, each representing a different aspect of usability. Each factor includes a different number of subquestions, scored on a 5-point Likert scale, resulting in varying maximum scores (5, 10, or 15). Scores by factor (with reverse scoring; higher is better), were 7.7/10 for RESPONSE ACCURACY (whether the system is accurate and does what the user expects), 14.1/15 for LIKEABILITY (whether the system is useful, the system is friendly, and it is clear how to send messages), 9.1/10 for COGNITIVE DEMAND (whether they felt confident or tense using the system), 7.6/10 for ANNOYANCE (whether the interaction was repetitive or boring), 7.3/10 for HABITABILITY (whether they always knew what to say to the system and knew what the system was doing), and 3.7/5 for SPEED (whether the interaction was fast). This indicates that the system is likable and makes participants feel confident, with room for improvement in habitability and speed.

6.2 Qualitative Coding

Next, we describe the results of our qualitative coding of participant interviews, reporting on GPTCoach's ability to adhere to each of our three design principles.

6.2.1 DP-1: Facilitation & Non-Prescriptiveness. We found that participants readily recognized facilitative and non-prescriptive qualities in their interaction with the system. When describing their overall experience with the chatbot, P10 said *"it sort of met me where I was at. [...] It first asked about sort of some contextual things before prescribing anything,"* while P5 shared *"I like how the system says, can you share? You know, it's not commanding."* Many of the participants recognized that a non-prescriptive approach ensures that advice is appropriate, such as P11: *"if you don't know what I do for work, or if you don't know what my daily schedule looks like and what things make me happy and what things I want to avoid, then it's like, you can't actually curate a good plan for me."* Many of the same participants appreciated having intentional time to reflect, such as P5 sharing that *"It made me really think about exercise and how positive it can be [...] When it brought up, who do you do this for? You know, what motivates you? It really, like, touched my heart a little bit."*

However, many participants were not accustomed to a computer adopting a facilitative role. P4 shared *"You can tell it's powered by a lot of positive recommendations and positive language and inclusive language. I think it is probably the most questions that I have ever been asked by a chatbot."* On the other hand, P2 felt like the chatbot was asking too many questions: *"it seems a little rapid fire, maybe just a lot of stuff. Yeah, so it's a little overwhelming."* This was likely impacted by GPT-4's tendency towards verbosity, where it would sometimes ask multiple questions in one message. Several participants simultaneously expressed that they appreciated the facilitative tone, while also wanting more prescriptive advice. P1 addressed this tension directly, sharing that *"when I face obstacles, and this is going to be the hard part, because here I am asking for someone to be like, 'Hey, you didn't do this,' like a taskmaster, but then, you know, in the moment, you're already feeling like, oh, I didn't do enough, [...] in that case, you would want this kindness."*

6.2.2 DP-2: Tailoring Advice and Integrating Context. Participants overwhelmingly expressed that the system accurately understood their preferences, constraints, and concerns, drawing attention to the personalized nature of its advice. For example, P7 shared that *"I really liked that it was accurate, that it was like my personal thing and not just abstract pictures before and after from the Internet of people who are not related,"* while P13 expressed that *"it felt like it was responding directly to the information I gave it, which is good because I think that's not always the case with the chatbots."* Participants built a mental model of the system's capabilities over time, learning that the system could take their context into account through explicit acknowledgments, such as P12 sharing *"There was acknowledgment, [...] with my shoulder, acknowledging previous injury."* These acknowledgments were particularly important as part of building trust in the system. For instance, P3 remarked early on that *"usually systems don't take [back pain] into account. so, like, I already don't have trust,"* later building more trust in the system as it appropriately acknowledged their concerns. P7 initially was

not sure if the chatbot would be able to understand their lack of motivation, noting that *"it's asking from the ideal world where robots don't have laziness, they have answers right away and they [are] always ready to work."*

In comparing our system to existing tools, participants highlighted its ability to capture additional, important context beyond sensor data. For instance, P10 shared that *"I like it just because I think of like, exercise and physical activity is not really like much of like a numbers game. [...] There's a lot more factors,"* while P12 noted that *"it was the kind of the personal touches [...] Different things that were not part of the Apple Fitness questions."* In reflecting on prior experiences with self-tracking, P4 shared that *"I've tried to use, like, MyFitnessPal from a food tracking standpoint. [...] While I can use the data, sometimes the emotional side of my brain gets more over than the logical side. So I would say, in comparison, I think this was easy."*

While GPTCoach consistently provided personalized and non-prescriptive guidance, participant impressions of advice specificity were more mixed. Some participants expressed a desire for more specific advice at the end of the conversation, particularly after having shared so much information about themselves. For example, P12 shared that they wanted *"maybe more detail, you know, kind of digging into some of my exercise data and maybe some of my feelings on how things are going."* Our chatbot would not always probe further for specific days and times, or which specific routines to do in each session. Several participants explicitly prompted the chatbot to be more specific, and many were satisfied with the resulting output but wanted the chatbot to have been more proactive. At the same time, other participants, particularly those who were older and/or less physically active, were highly satisfied with the higher-level advice they received. For example, P4 noted, *"When I see it written out like this, there's no reason I shouldn't do it."* P2 found the advice not only specific, but even more actionable than that of human coaches, stating that *"I think a lot of the conversations that I had with the [human] health coach were maybe not as specific [...] and that was one thing I kind of didn't like about and why I didn't continue that health coaching."* These mixed results suggest that participants have differing needs and expectations regarding the level of specificity in advice. While some participants appreciated granular details and structured plans, others preferred more flexible, high-level guidance. Crucially, none of the participants who wanted more specific advice expressed receiving any advice that was not useful or accurate.

Participants generally appreciated and were often impressed by the system's use of data. In its best moments, the system could fetch data at appropriate times and weave it into the conversation in ways that empowered the client. For example, in response to a participant sharing that they've been *"feeling down/depressed if I don't exercise,"* the system responded with *"Thank you for sharing that, [name]. It takes courage to discuss these frustrations, and I appreciate your openness. From your previous data, it seems you've been engaging in quite a diverse and healthy mix of activities in the past month. Look, you've managed a total of 35 workouts with varying duration! That's a wonderful achievement."* Showing interactive visualizations were also prompts for reflection, such as P12 remarking *"Apparently on the weekends, I don't do much. I sit on the couch and watch TV,"* or P14 appreciated the ability to ask questions about data, sharing that

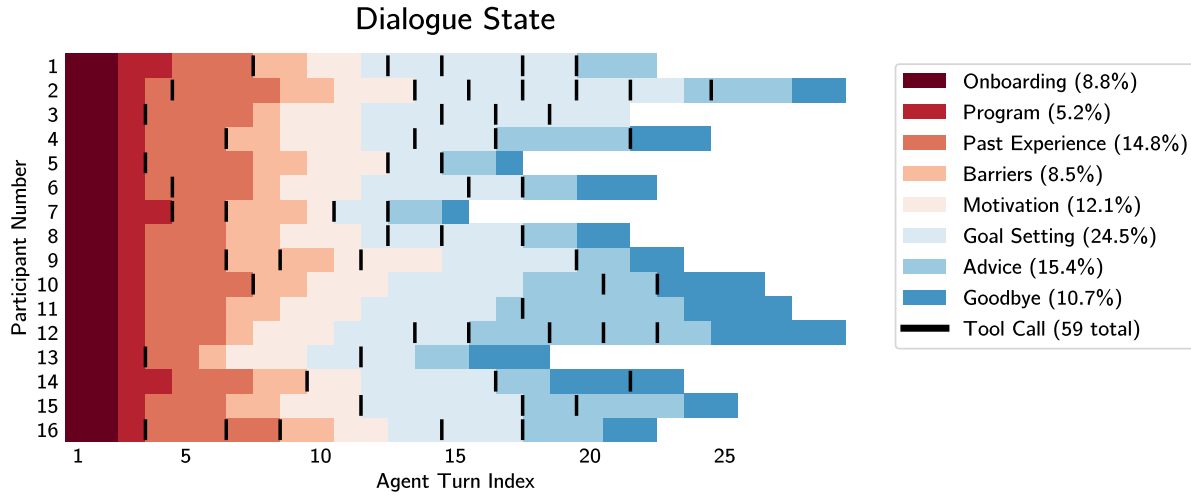


Figure 5: Progression of GPTCoach's Dialogue States By Turn Index. We find that GPTCoach adaptively allocates more conversational turns for gathering information about past experiences, barriers, motivation. GPTCoach allocates the most turns for the goal-setting state. Tools calls are appropriately called during past experience, goal setting, and advice.

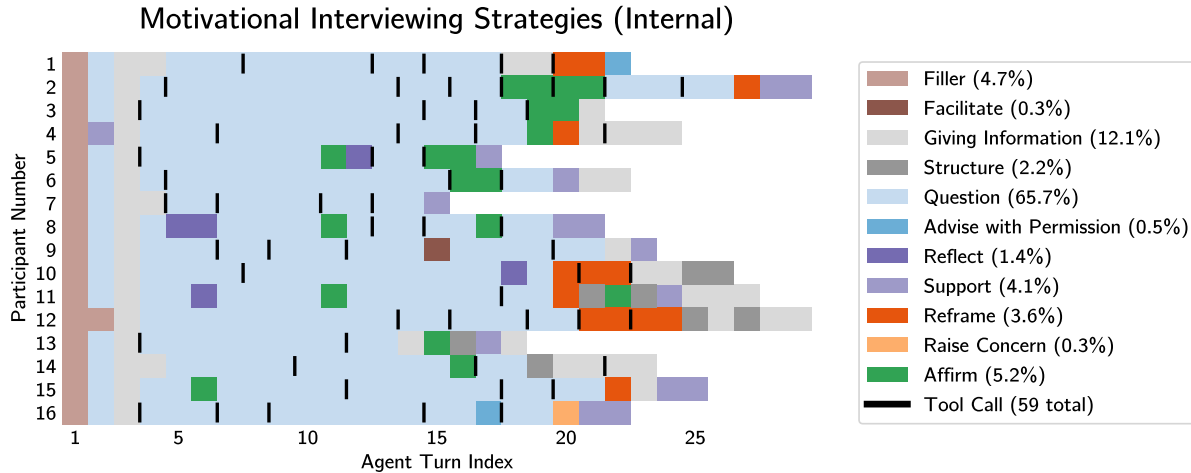


Figure 6: Progression of GPTCoach's Internal MI Strategies By Turn Index. We find that most of the conversation is spent asking questions and that QUESTION, REFLECT, and AFFIRM precede ADVISE WITH PERMISSION and GIVING INFORMATION. Questions and reflections preceding advice is more aligned with high-quality counselor behavior [109].

"I cannot ask any questions about the data that Apple Watch is, like, collecting, but I can ask questions from the chatbot."

However, in other moments, the system's use of data was more variable. Sometimes, the system would display a chart without acknowledging it in a follow-up message, leading to confusion. Other times, data discrepancies could sidetrack the conversation and break trust in the system, further complicated by the fact that the system tended to treat data as fact without asking the participant to confirm its accuracy or relevance. Interestingly, P8 told the chatbot that they did not use their phone to track steps, after which the system ignored their step count data. Moreover, several participants expected more of the system's data analysis capabilities than

was currently supported by our tool calls. These expectations may have been shaped by general impressions of computers' advantages over humans in data analysis, e.g., *"the real person, like therapist, they don't have time to read my data [...] it's a program that can read all of this like multiple data and the real person can't"* (P7). Some participants wanted more granular analysis of existing data, while others wanted additional data sources from other wearables not supported by our system.

6.2.3 DP-3: Supportive & Non-Judgmental Tone. Our strongest findings came from participants' impressions of the system's **positive**,

supportive, and empowering tone. Participants felt comfortable, supported, and safe when speaking with the chatbot: “I think the way the system answered, it didn’t make me feel like I was being judged.” (P2); “I think I’ve noticed that I feel some sort of, like, psychological safety with chatting with the chat bot.” (P3). Many compared the experience to interacting with a human coach, such as P6 noting that “I had a really positive experience. [...] it felt like I was, like, chatting with a human, honestly, or like a coach” (P6). Some participants shared that they preferred this interaction to prior interactions with humans. P7 told us that they felt more safe, noting that “and here it’s like, no gender, no body shape, nothing. [...] Yeah, comfortable, less judgmental. With the real person, you’re always comparing.” P5 shared that “I had a doctor who was, now, I was probably only, like, 20 pounds overweight at the time, and she told me I was morbidly obese, and she told me I needed to exercise. Like, I mean, it was, like, a ridiculous amount. [...] and I stopped going to her because I just felt so threatened.” Later, P5 shared that our system “doesn’t feel threatening at all.” For some, the positive tone was too much, such as P10 telling us “It is really good to be affirmed. But we’ve been affirmed quite a few times.” For others, the conversation felt decidedly neutral, “It did not emit any emotion out of me either way. [...] I was just having a conversation with a computer program as far as I can tell” (P12). Most importantly, none of the participants reported a negative or judgmental experience interacting with the chatbot.

6.3 Dialogue Analysis

Next, we report on a computational analysis of participants’ conversations with GPTCoach. We analyze GPTCoach’s internal dialogue states, MI strategies, and tool calls, as well as a comparison of GPTCoach’s message length.

6.3.1 Dialogue States: How does GPTCoach structure conversations?

As shown in Figure 5, we find that dialogue states follow a similar progression across all participants, with individual variation in the number of turns allocated to each state. The chatbot spends no more than 1-2 turns during onboarding or describing the program, but adaptively allocates more turns when gathering information about past experiences, barriers, or motivation. Overall, we notice that 28.5% of all messages were spent in goal setting, with an average of 5.6 agent responses during this state. This was followed by 15.4% of the total conversations in the advice state (an average of 3.5 messages) and 14.8% of the conversations in the past experiences state (an average of 3.4 agent responses). The messages between past experiences and motivation states account for a total of 35.4% of the total conversations. This behavior is more aligned with high-quality counselor behavior [109], while low-quality counselors focus on problem-solving before collecting sufficient context from the client [26, 34].

We also see that the chatbot almost exclusively performs tool calls between past experience and advice dialogue states. Of the 59 total tool calls across all users, 35.6% of the tool calls were in the goal setting state, 30.5% in the advice state, and 13.6% in the past experience state. This indicates that tool calls are mostly called at appropriate conversation states: collecting information about users’ past experiences, grounding goal setting in concrete historical baselines, or giving tailored advice.

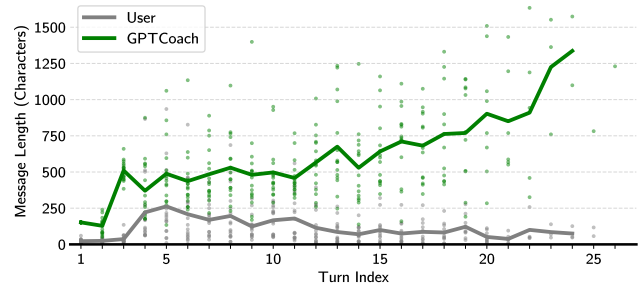


Figure 7: Comparing the Length of GPTCoach and User Responses. We find that GPTCoach’s responses are 2-3+ times as long as the user’s responses, suggesting that the system could be improved with shorter responses.

6.3.2 Response Length: How long are GPTCoach’s responses? We analyze the chatbot’s utterance length and compare them to the user’s utterance length. As seen in Figure 7, agent responses are 2-3 times the length of user responses, which increases as the conversation advances. Despite prompts to keep responses brief, the chatbot struggled to limit its verbosity, which is indicative of low-quality therapist behavior [109] and points to areas for future work.

6.3.3 MI Strategies: How does the chatbot use motivational interviewing strategies? We first look at the distribution of *internal* MISC strategies that the chatbot uses to ground its responses in the MI chain. As shown in Figure 6, most of GPTCoach’s responses are *QUESTIONS*, covering 65.7% of the entire conversation across all participants. This was followed by 12.1% of chatbot responses grounded in *GIVING INFORMATION* and 5.2% in *AFFIRM*. We find that *QUESTION*, *REFLECT*, and *AFFIRM* generally precede *ADVISE WITH PERMISSION* and *GIVING INFORMATION*, which is more aligned with high-quality counselor behavior [109].

6.4 Motivational Interviewing Coding: How well does the chatbot adhere to Motivational Interviewing?

We next examine the distribution of *external* MITI behavior codes present in GPTCoach’s utterances, as coded by human MI experts. On average, each agent response contained 2.3 different external strategies. As shown in Figure 8, the most frequent strategies are *QUESTION* (32.9% of all codes), *GIVING INFORMATION* (31.1%), and *AFFIRM* (13.6%). However, despite our effort to discourage unsolicited advice, we also encounter undesirable, MI-inconsistent codes like *PERSUADE* across 6.7% of all codes. MI-inconsistent codes tend to occur later in the conversation, as the agent shifts from asking questions to providing information and giving advice, some of which can be unsolicited. Aggregating these strategies into MI consistent (*AFFIRM*, *EMPHASIZE AUTONOMY*, *PERSUADE WITH PERMISSION*, *QUESTION*, *SEEKING COLLABORATION*, *SIMPLE REFLECTION*, *COMPLEX REFLECTION*), inconsistent (*PERSUADE*, *CONFRONT*), and neutral codes (*GIVING INFORMATION*), we see in Figure 9 that only 6.7% of total codes corresponding to our chatbot’s responses are inconsistent with MI principles.

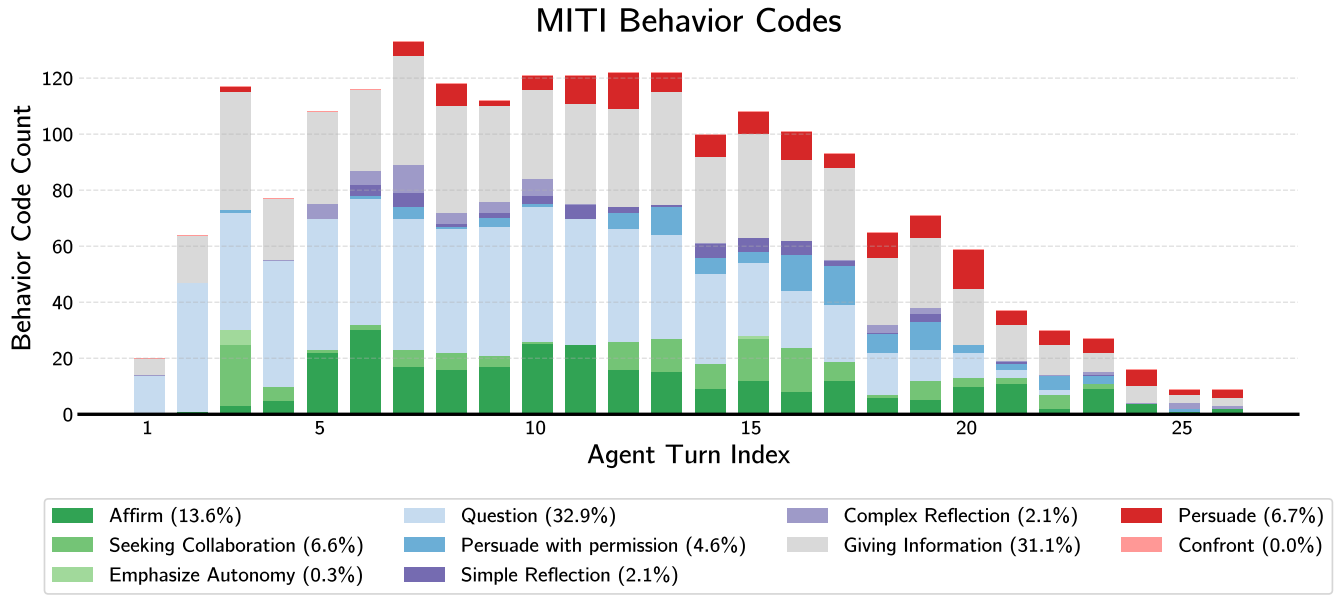


Figure 8: Distribution of GPTCoach’s External MITI Behavior Codes by Turn Index. We find that GPTCoach most frequently engages in AFFIRM, QUESTION, and GIVING INFORMATION, which is aligned with MI. However, GPTCoach still engages in ADVISE WITHOUT PERMISSION, which is not aligned with MI. MITI behavior code definitions are provided in Appendix C Table 12.

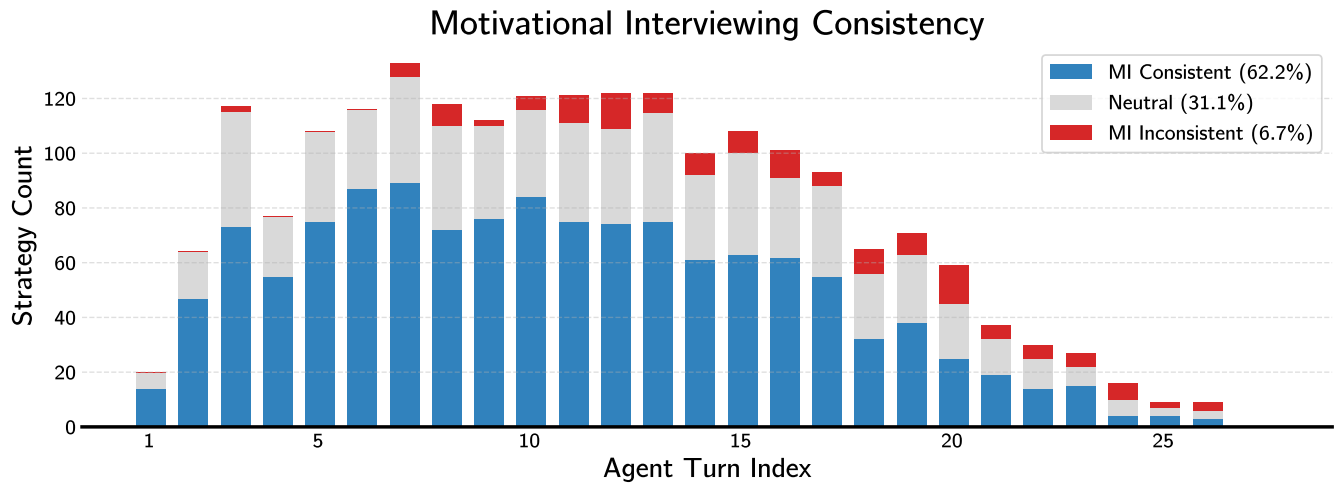
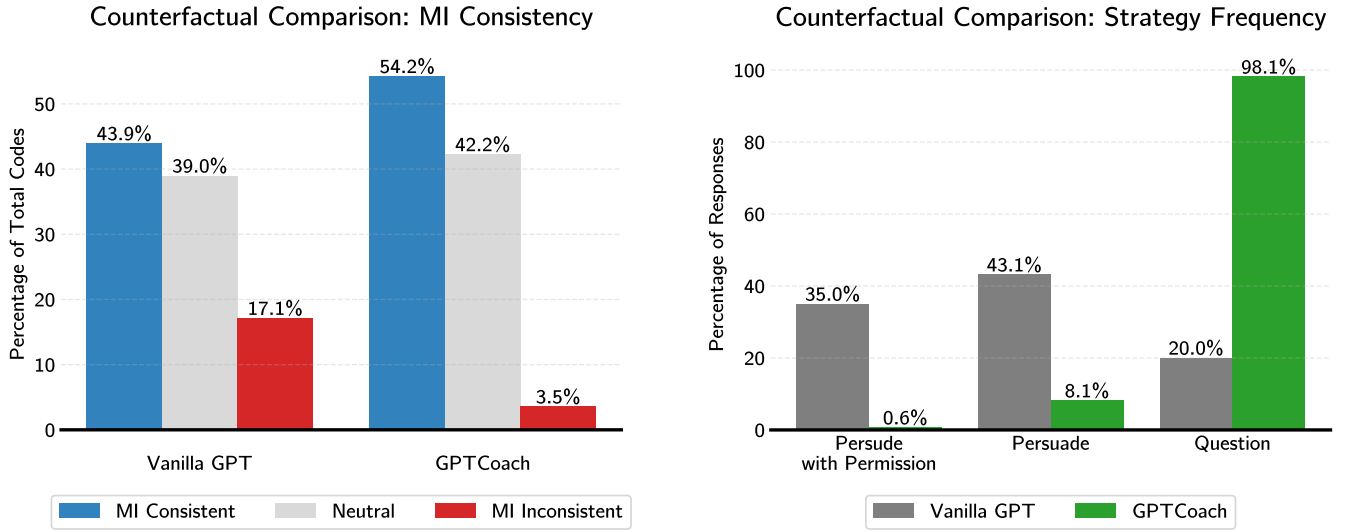


Figure 9: GPTCoach’s MI Consistency by Turn Index. This figure aggregates MITI behavior codes from Fig. 8 into MI-consistent, MI-inconsistent, and neutral codes [89]. We find that only 6.7% of GPTCoach’s total codes are inconsistent with MI principles.

6.4.1 Qualitative Feedback from MITI Coders. In addition to behavior coding, we also received qualitative feedback from the coding team through video-call discussions and a final feedback document. The coders noted that GPTCoach demonstrated a respectful and non-confrontational approach, emphasizing client strengths and autonomy rather than focusing on their difficulties, which is in alignment with MI principles and fosters a more open relationship. However, they also highlighted several areas for improvement. One experienced coder remarked that GPTCoach’s MITI scores were

similar to those of human counselors in training, who grasp basic building blocks but have yet to master MI. Specifically, GPTCoach tended to offer motivational statements or advice in response to “change talk” (client statements in favor of behavior change) rather than leveraging these moments to further elicit client motivation and explore the benefits of change. While some good reflections were present, they were underutilized, with GPTCoach often defaulting to providing information or asking questions instead of letting the client “hear their own thoughts and emotions in the



(a) Comparing the Overall Frequency of MI-consistent, MI-inconsistent, and Neutral Strategies in Our Counterfactual Analysis. Percentages represent the proportion of strategies that are MI-consistent/MI-inconsistent/neutral strategies among all MI strategy codes for a given agent. We find that GPTCoach’s responses contain more MI-consistent and fewer MI-inconsistent responses than vanilla GPT.

(b) Comparing the Frequency of MI Strategies by Response in Our Counterfactual Analysis. These three strategies showed the largest differences in frequency across the two agents. Percentages represent the proportion messages that contain a given MI strategy across all of a given agent’s counterfactual responses. We find that vanilla GPT uses persuasion (i.e., advice or solutions) far more frequently (and often without permission), while GPTCoach gives much less advice and asks far more questions.

Figure 10: Counterfactual Comparison Between GPTCoach and Vanilla GPT-4. The vanilla GPT condition includes the same system prompt, but with all prompt chains ablated. Illustrative examples are provided in Appendix D.1.4.

voice of another.” Skilled MI practitioners would use more reflections (particularly complex reflections), give less information, and avoid persuasion or advice altogether. Lastly, as confirmed by our analyses (Section 6.3.2), GPTCoach tended to speak more than the client, which is contrary to MI’s client-centric approach.

6.5 Counterfactual Analysis: How does GPTCoach compare to vanilla GPT-4?

We report on our counterfactual analysis comparing GPTCoach to GPT-4 using only the system prompt, with all prompt chains removed. As shown in Figure 10a, 54.2% of GPTCoach’s codes are MI-consistent, compared to 43.9% of vanilla GPT-4’s codes. Meanwhile, 3.5% of GPTCoach’s codes are MI-inconsistent, compared to 17.1% of vanilla GPT-4’s codes. We found that the top three codes driving this difference in MI-consistency are PERSUADE WITH PERMISSION, PERSUADE (without permission), and QUESTION, matching our intuitions. In Figure 10b, we plot the percentage of messages that contain each of these codes. We find that GPT-4 contains PERSUADE in 43.1% of its responses, compared to 8.1% of GPTCoach’s responses. Meanwhile, GPTCoach contains QUESTION in 98.1% of its responses, compared to 20.0% of vanilla GPT-4’s. Lastly, we also find that vanilla GPT-4 contains PERSUADE WITH PERMISSION in 35.0% of its responses, compared to 0.6% of GPTCoach’s replies. Vanilla GPT-4 has a stronger bias towards persuasion and giving

advice, frequently without permission, while GPTCoach is more inclined to ask questions before jumping to giving advice.

6.6 LLM-Based Motivational Interviewing Coding

We briefly report on our LLM-based coding results, which leveraged Chiu et al.’s [26] prompt-based method for evaluating LLM psychotherapists and used the MISC coding scheme [93]. While there are differences between the MITI and MISC codeset, our LLM-based coding results are highly analogous to our human coding results, indicating that GPTCoach is largely consistent with MI, with room for improvement in advanced MI skills. We find that 54.1% of codes are MI-consistent, 30.0% are neutral, and 15.9% are MI-inconsistent. The most commonly used strategies are AFFIRM (18.8%), OPEN QUESTION (18.0%), and GIVING INFORMATION (16.4%). In the counterfactual analysis using LLM-based coding, 62.6% of GPTCoach’s codes are MI-consistent, compared to 55.0% of vanilla GPT-4’s codes. Meanwhile, 14.5% of GPTCoach’s codes are MI-inconsistent, compared to 22.9% of vanilla GPT-4’s codes. Given the growing interest in LLM-based MI coding in HCI and NLP [90, 107–109, 117, 127, 128], these findings suggest that prompt-based methods show promise for automated MI coding. Full details on LLM-based coding are provided in Appendix D.2.

7 LIMITATIONS

A primary limitation of our current study is that we did not examine long term use of GPTCoach and thus could not evaluate its ability to sustain physical activity behavior change. As discussed in Section 4.2, we believe that many of our technical innovations and learnings will transfer to the multi-session setting. However, we emphasize that we opted for a single-session study design in this formative work such that researchers could supervise all interactions, thereby mitigating the risk of harm from unpredictable outputs. We believe that additional work is needed to ensure that GPTCoach is safe to use without researcher oversight and we discuss ethical considerations and possible risk mitigation strategies in Section 8.3.

We encountered several challenges with current models that may have limited our chatbot’s efficacy, such as difficulties adhering to the coaching program, effectively employing motivational interviewing strategies, and using tool calls at the appropriate times. In Section 8.2, we discuss how this behavior may arise from instruction-tuning and reinforcement learning from human feedback (RLHF) [105]. We were able to address many of these challenges through our prompt chaining architecture, which greatly improved MI-consistency compared to a baseline GPT-4. However, qualitative feedback from MI experts indicated that the system still fell short of achieving the performance level of skilled human coaches and future work could explore how to further improve adherence to MI. Moreover, our final system was not always consistent in utilizing data proactively. It is possible that this could be improved with finetuning or with more sophisticated data analysis functionality, such as multi-agent architectures or code generation [87]. Despite these challenges in incorporating data, participants’ ratings of advice quality and personalization were not noticeably affected (see Section 6.1.1 and Figure 4). This could suggest that personalization based on qualitative context is more impactful to user experience than data-driven personalization. Alternatively, it may reflect the difficulty of achieving meaningful personalization based on quantitative data, mirroring challenges faced by human coaches in incorporating data [113, 114].

Lastly, we acknowledge limitations in the demographics of our participant pool. While we aimed for a diverse sample across age, gender, race/ethnicity, physical activity levels, and stages of behavior change—including older adults, individuals from minority groups, non-native English speakers, individuals with chronic health conditions, and neurodivergent individuals—our participant pool reflects some socioeconomic bias. Participants were all based in the US and required to own iPhones, as our system relied on Apple HealthKit. iPhones ownership is associated with higher income and education [59], and all participants had some post-secondary education. While we believe that LLM coaching can increase access to physical activity support, future work that critically engages with socioeconomic class is needed to ensure the technology meets its potential for benefit [70].

8 DISCUSSION

Our formative interviews and technology probe study demonstrate promising evidence that LLMs can collect rich qualitative context about people’s unique life circumstances and use this context to

personalize its physical activity support. Moreover, GPTCoach’s conversational flexibility allowed it to acquire this information in a non-judgmental, supportive manner, reflecting many of the positive qualities of human coaching. We now discuss our findings, including implications for mobile health interactions beyond conversation, considerations for extending GPTCoach to multi-session coaching, implications for LLM training and evaluation, as well as the ethics and privacy risks of LLM-based health coaching.

8.1 Implications for Future Mobile Health Systems

While conversational interaction proved highly effective for eliciting rich qualitative information, we also believe that LLM agents could make use of qualitative context to facilitate a range of interactions for mobile health applications beyond simple chat interfaces. For instance, a mixed-initiative [54] system might couple a chat interaction with ambient displays [36, 72, 96], adjusting the display in response to both quantitative sensor data and qualitative information acquired through conversation. Such a system might also allow users to set more flexible goals, which could be adapted and renegotiated as users progress or encounter new obstacles [95]. By integrating push notifications and calendar entries, an agent could tailor activity plans, schedule reminders, and readjust plans based on natural language input (e.g., *“I can’t go on a run this Thursday because I need to pick up my kids”*). In fact, many participants suggested several such additions, like personalized reminders, scheduling support, guidance during life changes or injuries, and real-time assistance. These interactions could be implemented by seeding a controller LLM agent with qualitative context and augmenting its capabilities through tools (e.g., functions that control UI elements or schedule push notifications).

Participants also requested different personas, such as a regimented ‘trainer’ when they needed an extra push and an empathetic ‘counselor’ when they were feeling down. Such a persona could enhance engagement and adherence through narrative [96] or be represented as an embodied agent [17, 64] to increase the emotional and relational connection. Participants had mixed opinions on whether users should choose the personas or have the agent adapt, raising interesting open questions for design. Moreover, it is unclear whether increased emotional connection with an LLM-based coach should be encouraged, as this might lead to a dependence on the coach, conflicting with client empowerment and self-efficacy.

While our study did not incorporate any nudging or just-in-time adaptive interventions [97], the qualitative context elicited during an onboarding conversation could also help drive adaptive experimentation algorithms. For instance, an LLM agent could help the user set a specific and measurable behavior change goal, which could in turn be optimized by an adaptive experimentation algorithm [82]. Qualitative information about a user’s preferences, abilities, and constraints could help warm-start an algorithm with better priors over optimal intervention timing and content. When new obstacles arise or life circumstances change, LLMs could help detect a distribution shift to better adapt the underlying learning algorithm.

A natural extension of our current system would provide users with continuous access to GPTCoach over multiple coaching sessions. After the onboarding conversation, the Active Choices program consists of several follow-up conversations, scheduled every few weeks. Our prompt chaining architecture could be adapted to conduct follow-up conversations using modified dialogue state prompts. Qualitative feedback elicited in follow-up sessions could be used by the agent to adjust exercise plans, provide strategies for overcoming barriers, make adjustments to the client’s goals, and provide encouragement. When coupled with additional features described above (e.g., push notifications, calendar entries, and ambient displays), the agent could also serve as an accountability tool, scheduling reminders for planned workouts or helping clients monitor progress towards their goal.

8.2 Implications for LLM Training & Evaluation

Instruction-tuning and reinforcement learning from human feedback (RLHF) [105] optimize models for single-turn question answering, which prior work suggests can bias the model’s behavior towards problem-solving and advice-giving [26, 118]. While this behavior is sensible for “helpful and harmless” [11, 105] assistants, it runs contrary to foundational principles of frameworks like motivational interviewing [89]—conversations that empower clients to change are not served by problem-solving and advice-giving. Instruction-tuning also biases the model to call tools only when prompted (e.g., “*visualize last month’s step count*”), rather than proactively incorporating relevant data.

It is possible that steerability towards facilitative behaviors, as well as effective integration of context, will improve with model scale, particularly in light of a growing interest in multi-turn, information seeking objectives [8, 136]. With advancements in multimodal learning, future foundation models for sensor data [2] might jointly encode sensor data and text, enabling multimodal understanding without representing sensor data as text. However, as argued by Ma et al. [80] in the domain of LGBTQ+ mental health support, task-specific models may be more effective for handling sensitive subjects rather than repurposing general foundation models. This necessitates large-scale training and evaluation datasets for health behavior change, which are time-consuming to create, but could enable rigorous, evidence-based health coaching models. While one could feasibly collect a dataset of existing health coaching conversations, care is needed because human and automated coaching differ significantly and it may not be ideal to mirror human coaching exactly [89]. Such datasets could also include rigorous evaluations of ethics and safety risks, which we discuss in the following section.

8.3 Risks of LLM-Based Health Coaching

While LLMs present several promising opportunities for health coaching, these new capabilities also introduce new risks that must be addressed for the technology to realize its potential for benefit. Below, we discuss several of these risks along with possible mitigating strategies. We focus on ethical concerns and risks that are specific to LLM-based health coaching, not general behavior change or mobile health applications.

8.3.1 Privacy. The collection and use of personal health data raises significant privacy implications. Regarding technical security measures in our system, we used HIPAA-compliant storage with Google Firebase, encryption using HTTPS, and privacy access controls in our iOS app (Appendix B.1). Our system leveraged the OpenAI API, which does not use API data for training and does not retain data after 30 days². Although participants consented to sharing their health data, a preferable alternative would be to use self-hosted or on-device LLMs. Current on-device models were not performant enough at the time of our study, but their performance may improve with growing interest in small models and on-device inference [3, 6, 47].

We take a human-centric stance towards privacy, arguing that technical data security measures are necessary but insufficient to ensure privacy. For example, a user may consent to collect body temperature but be unaware this data can be used to expose pregnancy status. While our system did provide access controls, it provided little support for understanding the benefits and risks of sharing various data sources. In future work, an LLM agent might offer the potential to educate users about privacy decisions. On the other hand, the use of a conversational agent may unintentionally influence users to disclose more information than intended.

8.3.2 Bias & Equity. LLMs have documented biases that can negatively impact marginalized groups [19, 71, 126], such as perpetuating harmful stereotypes, performing poorly with certain subgroups, or withholding empathy in response to identity disclosures [40]. Many of our participants readily shared aspects of their identities or health conditions with our chatbot. Researchers with training in computing ethics supervised all interactions with GPTCoach and analyzed study transcripts, while participants were asked about negative aspects of their interaction in post-study interviews. We did not encounter nor did participants report any instances of bias or potentially harmful responses, though we were limited by our small sample size. A study at larger scale and/or without researcher supervision would likely require systematic redteaming [46]. Our study provided insights into factors that such a systematic investigation should account for, including (but not limited to) gender, race and ethnicity, access to resources (e.g., financial barriers or lack of housing), motor impairments and disability, neurodivergence, and cultural differences. In addition, while GPTCoach focuses on providing support for physical activity only, physical activity is often entangled with other sensitive topics such as nutrition and diet, weight loss, body image, eating disorders, mental health, substance use, sleep disorders, and medical concerns. Though these topics can be considered out of scope, an LLM agent should be able to acknowledge the concerns, point the client to relevant resources, and gently redirect the conversation back to physical activity without causing harm. GPTCoach exhibited preliminary evidence of being able to appropriately navigate conversations related to weight loss, nutrition, and mental health, but further systematic validation of our system’s robustness is warranted.

While no method can guarantee safe and unbiased outputs with certainty, we are encouraged by recent work on redteaming, guard models, and evaluation benchmarks [23, 25, 58, 85, 135]. We believe that many of these methods can also be leveraged for risk mitigation

²<https://openai.com/enterprise-privacy>

in the domain of health coaching. In addition, we draw particular attention to LLM-based mental health support, which has received considerable attention in the literature and shares many of the same risks as health coaching [41].

8.3.3 Hallucination & Factual Accuracy. Even the most performant language models are known to confidently “hallucinate” false information [84, 126]. This is of utmost concern when offering health-related advice. While our model was based on an evidence-based coaching program that did include guidelines for giving advice to common barriers, it did not have access to external knowledge bases via retrieval [73, 120]. We encountered two instances of the model justifying its responses with reputable sources (the Center for Disease Control and American College of Sports Medicine) that were not included in its prompts. While both instances were innocuous and happened to contain factually accurate information that matched the source, we find this behavior concerning: not only can guidelines change, but it points to a risk of hallucinating reputable sources for factually inaccurate information.

One popular mitigation strategy for hallucination is retrieval augmented generation [120]. For example, Merrill et al. [87] provide their PHIA agent with web search to retrieve relevant health information from reliable sources. As with bias, we also believe that a study conducted at larger scale and/or without researcher oversight should include systematic evaluations of factual accuracy and evidence-based advice. One such example is described by Cosentino et al. [37], who evaluated their PH-LLM model on the National Strength and Conditioning Association’s Certified Strength and Conditioning Specialists exam and received a passing score (88%). Beyond static benchmarks, it may also be necessary to evaluate a model’s ability to provide accurate information in the context of a personalized health coaching conversation.

Future work on GPTCoach will be guided by the risk areas identified above, with a focus on implementing effective mitigation strategies for privacy, bias, and factual accuracy, such as enhanced data security measures, auditing and redteaming for bias and factual accuracy, or technical approaches such as finetuning or retrieval-augmented generation. This is of particular importance for studies in which participants interact with an LLM agent without researcher oversight.

9 CONCLUSION

This work draws inspiration from health coaching to explore the potential for personalized, LLM-based physical activity support. Through formative interviews with 12 health professionals and 10 potential recipients of health coaching, we identify three key design principles for LLM-based health coaching that emphasize the importance of a non-prescriptive approach, the integration of both qualitative and quantitative context, and a non-judgmental tone. These principles center the client’s agency and motivation, establishing an alternative to the question-answering and advice-giving objectives common in prior work and commercial products. Through the design and evaluation of GPTCoach in a lab study with 16 participants, we demonstrate that LLMs can effectively implement the onboarding conversation of an evidence-based physical activity coaching program. We find promising evidence that participants felt comfortable sharing concerns with and supported

by GPTCoach, as well as that GPTCoach’s advice was personalized and actionable. By integrating qualitative context from conversational interaction with quantitative data from wearable devices, GPTCoach represents a promising step towards more personalized and effective mobile health interventions. We discuss our work’s implications for future mobile health applications, how GPTCoach might be extended to multi-session coaching, implications for LLM training and evaluation, as well as the risks of LLM-based health coaching.

ACKNOWLEDGMENTS

We thank the Stanford Institute for Human-Centered Artificial Intelligence (HAI) for partial funding support and the OpenAI researcher access program for providing API credits to support this research. We extend our heartfelt gratitude to the Health Equity Action Research and Technology Solutions (HEARTS) lab for sharing their expertise and resources, particularly Maria Ines Campero, Prof. Abby C. King, and the Active Choices facilitators. We thank John Brelsford, Danielle Sessler, and the HETI Maine coding team for their consultation and support for MITI coding. We thank all of the students who contributed to the design and development of our system, including Evelyn Hur, Bryant Jimenez, Niall Kehoe, Dhruv Naik, Evelyn Song, Romuald Thomas, Caroline Tran, and Anthony Xie, as well as the CS342 teaching staff for supervising our project team. Thank you to Andrea Cuadra, Ryan Louie, Omar Shaikh, and Danilo Symonette as well as the IxD research group for their continuous support and feedback throughout the project. We are grateful for Beleicia Bullock, Alan Cheng, Elizabeth Childs, Cyan DeVeaux, Jane E, Ralf Herbrich, Michelle Lam, Lindsay Popowski, Yujie Tao, Jordan Troutman, and Dora Zhao’s contributions in recruitment, pilot testing, and feedback on paper drafts. We thank Michael S. Bernstein for providing feedback on inter-annotator agreement statistics. Lastly, we thank all of our participants, without whom this work would not have been possible.

REFERENCES

- [1] Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2023. Conversational Health Agents: A Personalized LLM-Powered Agent Framework. *arXiv preprint arXiv:2310.02374* (2023).
- [2] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhayakumar Nallasamy, and Ian Shapiro. 2024. Large-scale Training of Foundation Models for Wearable Biosignals. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*.
- [3] Marah Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219* (2024).
- [4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [5] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 2023. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. *Journal of Medical Internet Research* 25 (2023), e40789.
- [6] Keivan Alizadeh, Seyed Iman Mirzadeh, Dmitry Belenko, S. Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2024. LLM in a flash: Efficient Large Language Model Inference with Limited Memory. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12562–12584. <https://doi.org/10.18653/v1/2024.acl-long.678>

- [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.
- [8] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. In *Conference on Language Modeling (COLM)*.
- [9] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. 2020. Activity River: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–9.
- [10] Francesco Baglivo, Luigi De Angelis, Virginia Casigiani, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, and Caterina Rizzo. 2023. Exploring the Possible Use of AI Chatbots in Public Health Education: Feasibility Study. *JMIR Medical Education* 9 (2023), e51421.
- [11] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
- [12] Albert Bandura. 1999. Social cognitive theory: An agentic perspective. *Asian Journal of Social Psychology* 2, 1 (1999), 21–41.
- [13] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the use of reflection in interactive system design. In *Proceedings of the 2014 Conference on Designing Interactive Systems*. 93–102.
- [14] Marit Bentvelzen, Jasmin Niess, Mikolaj P Woźniak, and Paweł W Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [15] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
- [16] Timothy W. Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of Biomedical Informatics* 39, 5 (2006), 556–571.
- [17] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (June 2005), 293–327. <https://doi.org/10.1145/1067860.1067867>
- [18] Timothy W. Bickmore, Daniel Schulman, and Candace L. Sidner. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of Biomedical Informatics* 44, 2 (2011), 183–197.
- [19] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- [20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2022. On the Opportunities and Risks of Foundation Models.
- [21] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- [22] Cynthia M Castro, Leslie A Pruitt, Matthew P Buman, and Abby C King. 2011. Physical activity program delivery by professionals versus volunteers: The TEAM randomized trial. *Health Psychology* 30, 3 (2011), 285.
- [23] Crystal Tin-Tin Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A Omiye, Akaash Kolluri, Akash Chaurasia, et al. 2024. Red Teaming Large Language Models in Medicine: Real-World Insights on Model Behavior. *medRxiv* (2024), 2024–04.
- [24] Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4130–4161. <https://doi.org/10.18653/v1/2024.acl-long.227>
- [25] Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs’ (Lack of) Multicultural Knowledge. *arXiv preprint arXiv:2404.06664* (2024).
- [26] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *arXiv preprint arXiv:2401.00820* (2024).
- [27] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Volda. 2022. Reflection in theory and reflection in practice: An exploration of the gaps in reflection support among personal informatics apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [28] Eun Kyoung Choe, Bongshin Lee, and m.c. schraefel. 2015. Characterizing visualization insights from quantified selfers’ personal data presentations. *IEEE Computer Graphics and Applications* 35, 4 (2015), 28–37.
- [29] 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.
- [30] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers’ practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1143–1152.
- [31] Chia-Fang Chung, Jonathan Cook, Elizabeth Bales, Jasmine Zia, and Sean A Munson. 2015. More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *Journal of Medical Internet Research* 17, 8 (2015), e203.
- [32] Domenic V Cicchetti and Sara A Sparrow. 1981. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency* 86, 2 (1981), 127–137.
- [33] Céline Clavel, Steve Whittaker, Anaïs Blacodon, and Jean-Claude Martin. 2018. WENner: A theoretically motivated approach for tailored coaching about physical activity. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1669–1675.
- [34] Jeff L Cochran and Nancy H Cochran. 2015. *The heart of counseling: Counseling skills through therapeutic relationships*. Routledge.
- [35] Geoffrey L Cohen and David K Sherman. 2014. The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology* 65 (2014), 333–371.
- [36] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: A field trial of ubit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1797–1806.
- [37] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. 2024. Towards a Personal Health Large Language Model. *arXiv preprint arXiv:2406.06474* (2024).
- [38] Cora L Craig, Alison I Marshall, Michael Sjöström, Adrian E Bauman, Michael L Booth, Barbara E Ainsworth, Michael Pratt, ULF Ekelund, Agneta Yngve, James F Sallis, et al. 2003. International physical activity questionnaire: 12-country reliability and validity. *Medicine & Science in Sports & Exercise* 35, 8 (2003), 1381–1395.
- [39] Alia J Crum and Ellen J Langer. 2007. Mind-set matters: Exercise and the placebo effect. *Psychological Science* 18, 2 (2007), 165–171.
- [40] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*.
- [41] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693* (2023).
- [42] Zachary Enghardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai “Orson” Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2, Article 56 (May 2024), 25 pages. <https://doi.org/10.1145/3659604>
- [43] Daniel A Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Quier Chen, et al. 2020. Mapping and taking stock of the personal informatics literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–38.
- [44] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Patricia Maes. [n.d.]. *PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models*. In *IEEE-EMBS International Conference on Biomedical and Health Informatics*.
- [45] Centers for Disease Control and Prevention. 2022. Physical Activity. <https://www.cdc.gov/physicalactivity/index.html>
- [46] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

- [47] Google. 2024. Gemma - A Family of Lightweight, State-of-the Art Open Models from Google. <https://ai.google.dev/gemma>
- [48] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 943–952.
- [49] Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8, 1 (2012), 23.
- [50] Health Level Seven International. 2023. HL7 FHIR Release 4. <https://www.hl7.org/fhir/>.
- [51] Narayan Hegde, Madhurima Vardhan, Deepak Nathani, Emily Rosenzweig, Cathy Speed, Alan Karthikesalingam, and Martin Seneviratne. 2024. Infusing behavior science into large language models for activity coaching. *PLOS Digital Health* 3, 4 (2024), e0000431.
- [52] Jennifer L Hicks, Melissa A Boswell, Tim Althoff, Alia J Crum, Joy P Ku, James A Landay, Paula ML Moya, Elizabeth L Murnane, Michael P Snyder, Abby C King, et al. 2023. Leveraging mobile technology for public health promotion: A multidisciplinary perspective. *Annual Review of Public Health* 44 (2023), 131–150.
- [53] Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3–4 (2000), 287–303.
- [54] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [55] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the Helper: Supporting Peer Counselors via AI-empowered Practice and Feedback. *arXiv preprint arXiv:2305.08982* (2023).
- [56] Dandan Huang, Melanie Tory, and Lyn Bartram. 2016. A Field Study of On-Calendar Visualizations. In *Proceedings of the 42nd Graphics Interface Conference*. 13–20.
- [57] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: Inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 17–24.
- [58] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. *arXiv preprint arXiv:2312.06674* (2023).
- [59] Maral Jamalova. 2019. The comparative study of the relationship between smartphone choice and socio-economic indicators. *International Journal of Marketing Studies* 11, 11 (2019), 10–5539.
- [60] Daoyan Jin, Hallgeir Halvari, Natalia Maehle, and Anja H Olafsen. 2022. Self-tracking behaviour in physical activity: A systematic review of drivers and outcomes of fitness tracking. *Behaviour & Information Technology* 41, 2 (2022), 242–261.
- [61] Bart A Kamphorst. 2017. E-coaching systems: What they are, and what they aren't. *Personal and Ubiquitous Computing* 21, 4 (2017), 625–632.
- [62] Elisabeth T Kersten-van Dijk, Joyce HDM Westerink, Femke Beute, and Wijnand A IJsselstein. 2017. Personal informatics, self-insight, and behavior change: A critical review of current literature. *Human-Computer Interaction* 32, 5–6 (2017), 268–296.
- [63] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. In *Proceedings of the Fifth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 248)*, Tom Pollard, Edward Choi, Pankhuri Singhal, Michael Hughes, Elena Sizikova, Bobak Mortazavi, Irene Chen, Fei Wang, Tasmie Sarker, Matthew McDermott, and Marzyeh Ghassemi (Eds.). PMLR, 522–539. <https://proceedings.mlr.press/v248/kim24b.html>
- [64] Abby C King, Maria Ines Campero, Jylana L Sheats, Cynthia M Castro Sweet, Michelle E Hauser, Dulce Garcia, Aldo Chazaro, German Blanco, Jorge Banda, David K Ahn, et al. 2020. Effects of counseling by peer human advisors vs computers to increase walking in underserved populations: The COMPASS randomized clinical trial. *JAMA Internal Medicine* 180, 11 (2020), 1481–1490.
- [65] Abby C King, Robert Friedman, Bess Marcus, Cynthia Castro, Melissa Napolitano, David Ahn, and Lawrence Baker. 2007. Ongoing physical activity advice by humans versus computers: The Community Health Advice by Telephone (CHAT) trial. *Health Psychology* 26, 6 (2007), 718.
- [66] Abby C. King, William L. Haskell, C. Barr Taylor, Robert DeBusk, Cynthia M. Castro, Leslie A. Pruitt, and Stanford Prevention Research Center staff. 2002. The Stanford Active Choices Program: Telephone-Assisted Counseling for Physical Activity. Stanford Health Promotion Resource Center, Stanford Prevention Research Center, Stanford University.
- [67] Abby C King, Eric B Hekler, Cynthia M Castro, Matthew P Buman, Bess H Marcus, Robert H Friedman, and Melissa A Napolitano. 2014. Exercise advice by humans versus computers: Maintenance effects at 18 months. *Health Psychology* 33, 2 (2014), 192.
- [68] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3063–3072. <https://doi.org/10.1145/1978942.1979396>
- [69] 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.
- [70] Whitney-Jocelyn Kouaho and Daniel A Epstein. 2024. Socioeconomic Class in Physical Activity Wearables Research and Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [71] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3299–3321. <https://doi.org/10.18653/v1/2023.eacl-main.241>
- [72] Nicholas Lane, Mashfiqui Mohammad, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2012. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th International ICST Conference on Pervasive Computing Technologies for Healthcare*.
- [73] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [74] 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.
- [75] Daniel Lieberman. 2021. *Exercised: Why something we never evolved to do is healthy and rewarding*. Vintage.
- [76] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *UbiComp 2006: Ubiquitous Computing: 8th International Conference, Orange County, CA, USA, September 17-21, 2006 Proceedings* 8. Springer, 261–278.
- [77] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
- [78] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. *arXiv preprint arXiv:2305.15525* (2023).
- [79] Tiffany Christina Luo, Adrian Aguilera, Courtney Rees Lyles, and Caroline Astrid Figueroa. 2021. Promoting physical activity through conversational agents: Mixed methods systematic review. *Journal of Medical Internet Research* 23, 9 (2021), e25486.
- [80] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, Article 872, 15 pages. <https://doi.org/10.1145/3613904.3642482>
- [81] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105.
- [82] Aishwarya Mandyam, Matthew Jörke, William Denton, Barbara E. Engelhardt, and Emma Brunskill. 2024. Adaptive Interventions with User-Defined Goals for Health Behavior Change. In *Proceedings of the Fifth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 248)*, Tom Pollard, Edward Choi, Pankhuri Singhal, Michael Hughes, Elena Sizikova, Bobak Mortazavi, Irene Chen, Fei Wang, Tasmie Sarker, Matthew McDermott, and Marzyeh Ghassemi (Eds.). PMLR, 597–618. <https://proceedings.mlr.press/v248/mandyam24a.html>
- [83] Bess H Marcus, Vanessa C Selby, Raymond S Niaura, and Joseph S Rossi. 1992. Self-efficacy and the stages of exercise behavior change. *Research Quarterly for Exercise and Sport* 63, 1 (1992), 60–66.
- [84] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>

- [85] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhae, Nathaniel Li, Steven Basart, Bo Li, et al. [n. d.]. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *Forty-first International Conference on Machine Learning*.
- [86] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: An intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 849–858.
- [87] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. 2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464* (2024).
- [88] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language Models Still Struggle to Zero-shot Reason about Time Series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 3512–3533. <https://doi.org/10.18653/v1/2024.findings-emnlp.201>
- [89] W.R. Miller and S. Rollnick. 2023. *Motivational Interviewing: Helping People Change and Grow* (4th ed.). Guilford Publications.
- [90] Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-Aware margin Ranking for Counselor Reflection Scoring in Motivational Interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 148–158. <https://doi.org/10.18653/v1/2022.emnlp-main.11>
- [91] Elliot G Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N Tobin, Patricia Davidson, Arlene M Smaldone, and Lena Mamykina. 2021. Automated vs. Human Health Coaching: Exploring Participant and Practitioner Experiences. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–37.
- [92] Julia S. Mollee, Anouk Middelweerd, Saskia J. te Velde, and Michel C. A. Klein. 2017. Evaluation of a personalized coaching system for physical activity: User appreciation and adherence. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (Barcelona, Spain) (PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 315–324. <https://doi.org/10.1145/3154862.3154933>
- [93] Theresa Moyers, Tim Martin, Delwyn Catley, Kari Jo Harris, and Jasjit S Ahluwalia. 2003. Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy* 31, 2 (2003), 177–184.
- [94] Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment* 65 (2016), 36–42.
- [95] Sean A Munson and Sunny Consolvo. 2012. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 25–32.
- [96] Elizabeth L Murnane, Xin Jiang, Anna Kong, Michelle Park, Weili Shi, Connor Soohoo, Luke Vink, Iris Xia, Xin Yu, John Yang-Sammataro, et al. 2020. Designing ambient narrative-based interfaces to reflect and motivate physical activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [97] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* (2018), 1–17.
- [98] Yoo Jung Oh, Jingwen Zhang, Min-Lin Fang, and Yoshimi Fukuoka. 2021. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *International Journal of Behavioral Nutrition and Physical Activity* 18 (2021), 1–25.
- [99] Jeanette M Olsen. 2014. Health coaching: A concept analysis. In *Nursing Forum*, Vol. 49. Wiley Online Library, 18–29.
- [100] Jeanette M Olsen and Bonnie J Nesbitt. 2010. Health Coaching to Improve Healthy Lifestyle Behaviors: An Integrative Review. *American Journal of Health Promotion* 25, 1 (2010), e1–e12.
- [101] Qi Chwen Ong, Chin-Siang Ang, Davidson Zun Yin Chee, Ashwini Lawate, Frederick Sundram, Mayank Dalakoti, Leonardo Pasalic, Daniel To, Tatiana Erlikh Fox, Iva Bojic, et al. 2024. Advancing health coaching: A comparative study of large language model and health coaches. *Artificial Intelligence in Medicine* 157 (2024), 103004.
- [102] ONVY. 2024. <https://www.onvy.health>.
- [103] World Health Organization. 2022. Physical Activity Fact Sheet. <https://www.who.int/news-room/fact-sheets/detail/physical-activity>
- [104] World Health Organization. 2024. S.A.R.A.H, a Smart AI Resource Assistant for Health. <https://www.who.int/campaigns/s-a-r-a-h>.
- [105] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [106] Soya Park and Chinmay Kulkarni. 2023. Thinking Assistants: LLM-Based Conversational Assistants that Help Users Think By Asking rather than Answering. *arXiv preprint arXiv:2312.06024* (2023).
- [107] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 42–51.
- [108] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1426–1435.
- [109] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? Learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 926–935.
- [110] James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American Journal of Health Promotion* 12, 1 (1997), 38–48.
- [111] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (2016), 1–17.
- [112] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996).
- [113] Heleen Rutjes, Martijn C Willemsen, Milou A Feijt, and Wijnand A IJsselstein. 2022. The Influence of Personal Health Data on the Health Coaching Process. *Frontiers in Big Data* 5 (2022), 678061.
- [114] Heleen Rutjes, Martijn C Willemsen, and Wijnand A IJsselstein. 2019. Beyond behavior: the coach’s perspective on technology in health coaching. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [115] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2023), 68539–68551.
- [116] Paul Schmiedmayer, Vishnu Ravi, and Oliver Aalami. 2024. *Spezi*. <https://doi.org/10.5281/zenodo.10482368>
- [117] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.
- [118] Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309* (2023).
- [119] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.
- [120] 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>
- [121] Ben Singh, Timothy Olds, Jacinta Brinsley, Dot Dumuid, Rosa Virgara, Lisa Matricciani, Amanda Watson, Kimberley Szeto, Emily Eglitis, Aaron Miatke, et al. 2023. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *npj Digital Medicine* 6, 1 (2023), 118.
- [122] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2696–2707.
- [123] Alice Thudt, Dominikus Baur, Samuel Huron, and Sheelagh Carpendale. 2015. Visual Mementos: Reflecting Memories with Personal Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 369–378.
- [124] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards Conversational Diagnostic AI. *arXiv preprint arXiv:2401.05654* (2024).
- [125] Corneel Vandelanotte, Stewart Trost, Danya Hodgetts, Tasadduq Imam, Mamunur Rashid, Quyen G To, and Carol Maher. 2023. Increasing physical activity using a just-in-time adaptive digital assistant supported by machine learning: A novel approach for hyper-personalised mHealth interventions. *Journal of Biomedical Informatics* 144 (2023), 104435.

- [126] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
- [127] Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3315–3330.
- [128] Anuradha Welivita and Pearl Pu. 2023. Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*. 5411–5432.
- [129] WHOOP. 2023. Introducing WHOOP Coach, Powered By OpenAI. <https://www.whoop.com/us/en/thelocker/introducing-whoop-coach-powered-by-openai/>.
- [130] Sara Wilcox, Marsha Dowda, Sarah F Griffin, Carol Rheame, Marcia G Ory, Laura Leviton, Abby C King, Andrea Dunn, David M Buchner, Terry Bazzarre, et al. 2006. Results of the first year of active for life: Translation of 2 evidence-based physical activity programs for older adults into community settings. *American Journal of Public Health* 96, 7 (2006), 1201–1209.
- [131] Sara Wilcox, Marsha Dowda, Laura C Leviton, Jenny Bartlett-Prescott, Terry Bazzarre, Kimberly Campbell-Voytal, Ruth Ann Carpenter, Cynthia M Castro, Diane Dowdy, Andrea L Dunn, et al. 2008. Active for life: Final results from the translation of two physical activity programs. *American Journal of Preventive Medicine* 35, 4 (2008), 340–351.
- [132] Ruth Q Wolever, Leigh Ann Simmons, Gary A Sforzo, Diana Dill, Miranda Kaye, Elizabeth M Bechard, Mary Elaine Southard, Mary Kennedy, Justine Vosloo, and Nancy Yang. 2013. A Systematic Review of the Literature on Health and Wellness Coaching: Defining a Key Behavioral Intervention in Healthcare. *Global Advances in Health and Medicine* 2, 4 (2013), 38–57.
- [133] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [134] JD Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2206–2220.
- [135] Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. 2024. AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies. *arXiv preprint arXiv:2407.17436* (2024).
- [136] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 62178–62209. <https://proceedings.mlr.press/v235/zhou24t.html>

A FORMATIVE STUDY DETAILS

We conducted a one-hour semi-structured interview with each participant. For health experts, we first asked a series of questions guided by the following structure:

- What does a typical day look like for you?
- Why do you do the work that you do?
- How did you approach your relationship with clients?
- What types of clients did you typically work with?
- What are some of the common challenges that your clients face?
- How do you help people overcome those barriers? Do you have any common strategies or techniques?
- How do you adapt your strategies to cater to different people's needs and circumstances?
- Are there any strategies you tried in the past but no longer use?
- Do you use any technology to assist you during your own exercise?
- Do your clients use any digital technologies?
- How do you think technology changes your relationship with clients?
- How do you feel about the growing role of technology within the health/wellness space?

For non-experts, we used the following set of questions to guide our interview:

- How would you define the term “physical activity”? What about “exercise”? What about “fitness”?
- What kinds of activities do you do for exercise? (if applicable)
- How many days per week do you exercise in a typical week? (if applicable)
- Where do you exercise? (if applicable)
- Do you try to get people to be active with you? (if applicable)
- Have your levels of physical activity changed over time?
- If possible, can you tell us about a time in your life when you were particularly active?
- If possible, can you tell us about a time in your life when you were particularly inactive?
- What helps you stay motivated?
- What are some challenges you face in being active? Have you tried anything to overcome these challenge?
- Are you currently interested in getting more physical activity? If so, do you have any goals?
- Do you use any technology to assist you during your exercise? Why?
- Were there any technologies you tried before, but no longer use? Why?

In the last portion of our interview, both groups of participants were asked to *“Imagine that you have access to an artificial intelligence chatbot that can help you improve your physical activity. The chatbot has access to information from a fitness tracker (such as an Apple Watch or Fitbit) as well as suggestions and strategies for improving your exercise. We’re going to be asking you questions about a number of different features this hypothetical technology might offer.”* We then asked participants to consider each of the following features, one at a time:

- Imagine that this chatbot is able to help you set effective goals
- Imagine that this chatbot is able to help keep you accountable towards those goals.
- Imagine that this chatbot is able to help you formulate a training plan for an upcoming event (e.g., a competition, or a wedding).
- Imagine that this chatbot is able to help you adjust your exercise routine in response to an injury.
- Imagine that this chatbot is able to help you visualize data from your fitness tracker and provide you with personalized insights.
- Imagine that this chatbot can adopt different personas, such as an empathetic wellbeing counselor or a regimented personal trainer.
- Imagine that this chatbot has access to your personal journal and offers insights into your mental wellbeing, personal relationships, and barriers to getting physical activity.

For each potential feature, we asked participants which aspects they liked, disliked, or had concerns about.

The research team synthesized these features through several research activities. First, we reviewed relevant literature on human health coaching and coaching manuals provided to us by our collaborators. We then individually prepared short storyboards grounded in the health coaching literature and our own explorations using GPT-4 to interpret our personal data. We presented our storyboards to an external group of researchers and used this feedback to synthesize a set of core interactions. These interactions were presented to our collaborators and in our research group for additional feedback.

B GPTCOACH: IMPLEMENTATION DETAILS

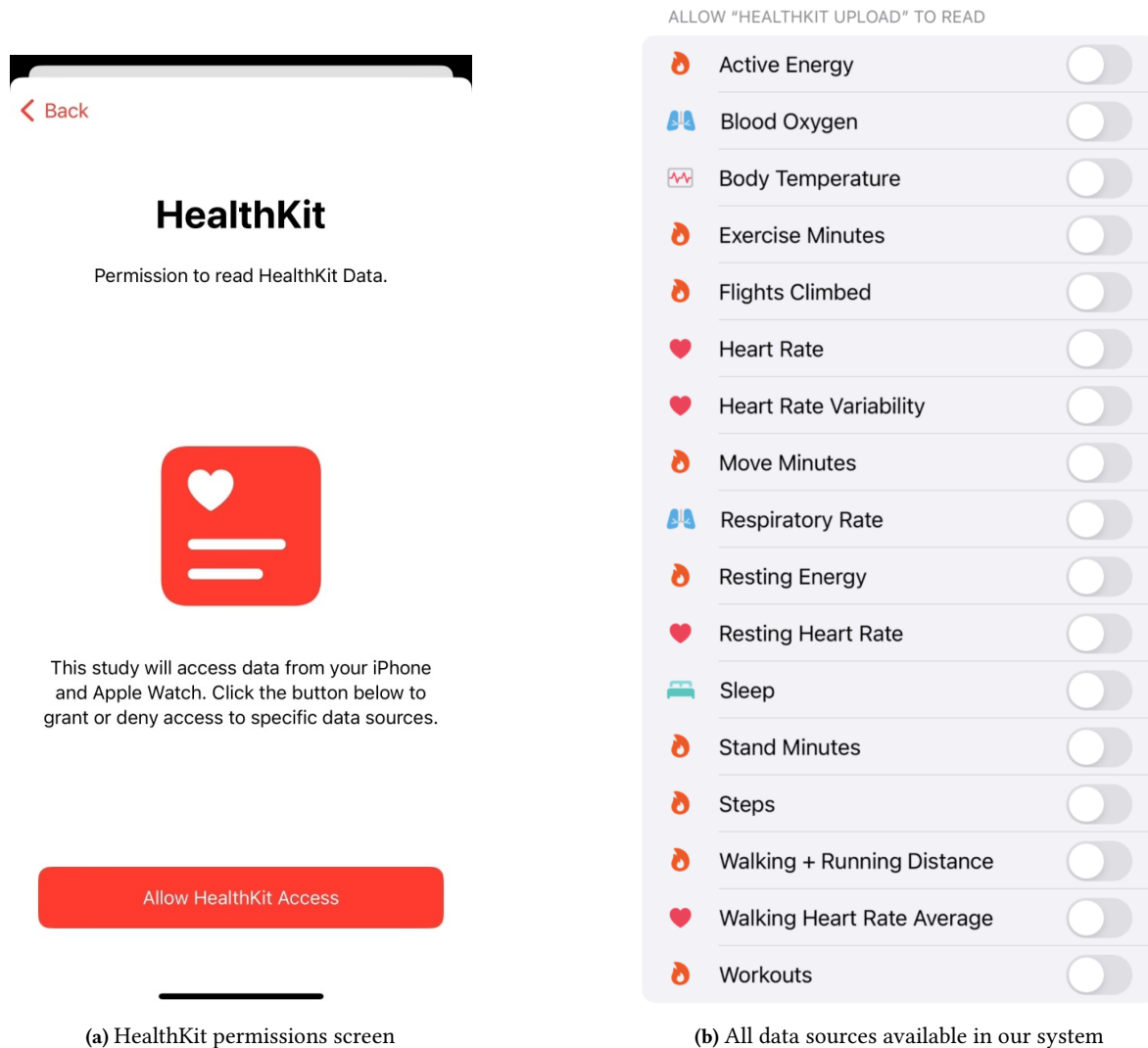


Figure 11: HealthKit permissions screen and data sources from our iOS application

B.1 iOS Application

To fetch users' historical data from the Apple HealthKit API and upload it to our Firestore database, we developed an iOS application using the Spezi open source framework [116]. Prior to installing our application, participants signed a consent form that informed them that they would be uploading three months of their health data and that they would be interacting with a chatbot system that has access to this data. Selected participants were also reminded of this in email communications prior to scheduling a study session. Participants were also informed that they could deny access to whichever data source they did not feel comfortable sharing.

As part of the app's onboarding, users are shown a permissions screen (Figure 11a) which requests access to their HealthKit data. Upon granting permissions to read HealthKit data, participants were shown the default iOS HealthKit permissions screen, which contains granular toggles for individual data sources (Figure 11b). Users who did not own an Apple Watch were able to share active energy, basal energy, flights climbed, step count, and walking + running distance. Participants who had other fitness trackers that sync with HealthKit (e.g., Oura or WHOOP) were also able to upload their data through our application.

B.2 Prompt Chains

In this section, we provide the structure for each of our prompt chains. LLMs are known to have poor performance when following instructions with long contexts. Given that instruction following is highest when relevant information is either at the beginning or end of the context [77], we re-iterate important task instructions relevant to the current prompt chain to the agent at the end of the context window as an assistant message. The general structure of our prompts thus includes: a 1) system prompt, 2) dialogue history, and 3) an agent prompt. Full prompt texts are all provided in Appendix E.

B.2.1 Dialogue State Chain. Upon receiving a new user message, the dialogue state chain first classifies whether or not to advance to the next dialogue state. This dialogue state classifier uses the following prompt structure:

Role	Prompt
System:	Dialogue State Classification Instructions (Fig. 15)
	– Dialogue History –
Agent Prompt:	Dialogue State Classifier Agent Instructions (Fig. 16)

Table 4: Prompt structure for dialogue state classification

B.2.2 Motivational Interviewing Chain. After deciding on the dialogue state, the motivational interviewing chain decides how to ground the model’s responses to the 11 Motivational Interviewing strategies. This motivational interviewing chain uses the following prompt structure:

Role	Prompt
System:	System Instructions (Fig. 18) + Dialogue State Prompt (Fig. 17) + Predict Strategy Instructions (Fig. 19) + MI Interviewing Strategies (Fig. 20)
	– Dialogue History –
Agent Prompt:	Strategy Prediction Agent Instructions (Fig. 21)

Table 5: Prompt structure for motivational interviewing strategy prediction.

B.2.3 Response Generation. After the MI strategy prediction, our prompt chain then uses this strategy to predict the response from the LLM. This response generation step uses the following prompt structure:

Role	Prompt
System:	System Instructions (Fig. 18) + Dialogue State Prompt (Fig. 17) + Generate Response Instructions (Fig. 22) + MI Interviewing Strategies (Fig. 20) + Few-Shot Tool Call Examples (Fig. 23)
	– Dialogue History –
Agent Prompt:	Response Generation Agent Instructions (Fig. 24)

Table 6: Prompt structure for response generation.

B.2.4 Tool Call Prediction. We introduce a tool call prediction prompt chain to improve the timeliness of fetching and visualizing a user’s health data. If the response generation step does not call a tool, using the output from the previous response generation step, we use an external LLM to predict whether tool call is appropriate to augment the conversation between the user and GPTCoach. This Tool Call Prediction chain uses the following prompt structure:

Role	Prompt
System:	System Instructions (Fig. 18) + Dialogue State Prompt (Fig. 17) + Tool Call Prediction Instructions (Fig. 25) + Few-Shot Tool Call Examples (Fig. 23)
	– Dialogue History + Agent Response –
Agent Prompt:	Tool Call Prediction Agent Instructions (Fig. 26)

Table 7: Prompt structure for tool call prediction.

B.2.5 Tool Call Generation. If the response generation step did not have a tool call and if the tool call prediction step determined that tool call was appropriate in the conversation, we introduced another agent to determine which tool call is appropriate in the conversation. This Tool Call Generation chain uses the following prompt structure:

Role	Prompt
System:	System Instructions (Fig. 18) + Dialogue State Prompt (Fig. 17) + Tool Call Generation Instructions (Fig. 27) + Few-Shot Tool Call Examples (Fig. 23)
	– Dialogue History + Agent Response –
Agent Prompt:	Tool Call Prediction Agent Instructions (Fig. 28)

Table 8: Prompt structure for tool call generation.

C EVALUATION STUDY DETAILS

In this section, we provide the survey measures for our evaluation study along with additional details for our automated motivational interviewing coding and counterfactual analyses. We also provide randomly sampled examples of our counterfactual analyses.

C.1 Survey Measures

C.1.1 User Experience & Quality of Advice. We asked participants the following survey items after interacting with GPTCoach. Each question was rated on a 5-point Likert scale, from 1: Strongly disagree to 5: Strongly agree.

The chatbot’s advice was actionable.
The chatbot’s advice was personalized.
The chatbot’s advice was generic.
I felt comfortable sharing my concerns with the chatbot.
I felt supported by the chatbot.
The chatbot made me feel capable of overcoming challenges.
The chatbot made me feel more motivated to change.
The chatbot asked me for my opinion about what activities I would like to do.
The chatbot understood my unique situation and concerns.
The chatbot gave me unsolicited advice.
The chatbot was empathetic.
The chatbot used my data in a way that was relevant.
The chatbot helped me identify obstacles to engaging in physical activity.
The chatbot helped me reflect on what motivates me to be physically active.
The chatbot helped make my own ideas about how to increase my physical activity more specific.
Interacting with the chatbot provided me with new insights about my physical activity.

Table 9: User Experience & Quality of Advice Questions

C.1.2 Subjective Assessment of Speech Systems Interfaces (SASSI). We measured usability using a subset of the Subjective Assessment of Speech System Interfaces [53]. We use the same subset as Mitchell et al. [91] with two additional questions from the habitability and speed factors. Each question was rated on a 5-point Likert scale, from 1: Strongly disagree to 5: Strongly agree.

Subscale	Question
RESPONSE ACCURACY	The system is accurate
RESPONSE ACCURACY	The system didn’t always do what I wanted
LIKEABILITY	The system is useful
LIKEABILITY	The system is friendly
LIKEABILITY	It is clear how to send messages to the system
COGNITIVE DEMAND	I felt confident using the system
COGNITIVE DEMAND	I felt tense using the system
ANNOYANCE	The interaction with the system is repetitive
ANNOYANCE	The interaction with the system is boring
HABITABILITY	I always knew what to say to the system
HABITABILITY	I was not always sure what the system was doing
SPEED	The interaction with the system is fast

Table 10: Usability Questions (SASSI [53])

C.2 Remote vs. In-Person Participants

In this section, we report demographics, usability, and advice quality conditioned on remote (11 participants) and in-person (5) participant subgroups. We did not observe notable differences across the two groups, both in our quantitative data (reported here) or in our qualitative feedback. In Table 11 below, we report demographics within each subgroup.

The SASSI [53] usability evaluation showed similar overall scores for Zoom (48.8/60, 81.3%) and In-Person (50.8/60, 84.7%) participants. Scores by factor for Zoom vs. In-Person were: Response Accuracy (7.6 vs. 7.8), Likeability (13.7 vs. 14.8), Cognitive Demand (8.9 vs. 9.6), Annoyance (7.5 vs. 7.6), Habitability (7.4 vs. 7.2), and Response Time (3.6 vs. 3.8). These results suggest that in-person participants reported marginally higher usability.

The advice quality evaluations also showed minimal differences between Zoom and In-Person participants. Scores by question for Zoom vs. In-Person were: Actionable (4.2 vs. 4.6), Personalized (4.5 vs. 4.6), Generic (2.5 vs. 2.6), Share Concerns (4.7 vs. 5.0), Supported (4.7 vs. 4.8), Capable (4.3 vs. 4.6), Motivated (3.9 vs. 4.2), Opinion Considered (4.8 vs. 4.2), Understood as Unique (4.4 vs. 4.6), Not Unsolicited (2.0 vs. 1.4), Empathetic (4.5 vs. 4.4), Relevant Data (3.9 vs. 3.8), Identify Obstacles (4.0 vs. 3.4), Reflect Motivation (4.3 vs. 4.4), Made Own Ideas Specific (4.1 vs. 4.2), and New Insights (4.3 vs. 3.2).

Age	<i>Zoom</i>	Mean: 39.4, Median: 33, SD: 16.0, Min: 21, Max: 71
	<i>In-Person</i>	Mean: 35.6, Median: 27, SD: 13.2, Min: 27, Max: 57
Gender	<i>Zoom</i>	Female: 6, Male: 5
	<i>In-Person</i>	Female: 4, Male: 1
Race/Ethnicity	<i>Zoom</i>	White: 8, East Asian: 1, Hispanic or Latino: 1, Middle Eastern: 1, Southeast Asian: 2, South Asian: 1
	<i>In-Person</i>	White: 3, Hispanic or Latino: 1, African-American or Black: 1
Education	<i>Zoom</i>	Associate: 2, Bachelor's: 4, Master's: 4, Doctorate: 1
	<i>In-Person</i>	Bachelor's: 2, Master's: 3
Stage of Change	<i>Zoom</i>	Precontemplation: 1, Contemplation: 6, Action: 2, Maintenance: 2
	<i>In-Person</i>	Contemplation: 2, Action: 1, Maintenance: 2
Level of Activity (IPAQ)	<i>Zoom</i>	Low: 4, Moderate: 4, High: 3
	<i>In-Person</i>	Low: 1, Moderate: 3, High: 1
AI Knowledge	<i>Zoom</i>	Basic: 7, Intermediate: 3, Advanced: 1
	<i>In-Person</i>	Novice: 1, Basic: 3, Advanced: 1

Table 11: Participant demographics by format (Zoom vs. In-Person) in the technology probe evaluation study ($N = 16$)

D MOTIVATIONAL INTERVIEWING CODING

In this section, we report on two motivational interviewing coding analysis. First, we report on our paper's primary analysis, in which human MI experts coded transcripts according to the MITI coding scheme. We then report on an analogous LLM-based coding analysis using the MISC coding scheme.

D.1 MITI Coding (Human Experts)

We hired trained coders to code transcripts according to the Motivational Interviewing Treatment Integrity (MITI) 4 Code [94]. We partnered with an agency that provides a wide range of MI training services, including MITI coding, and had previous experience coding chatbot transcripts. All of the coders were trained in MITI coding and had extensive coding experience.

D.1.1 Behavior Codes. In MITI coding, each “volley” (i.e., a turn or message) is partitioned into utterances and each utterance receives a unique behavior code. All MITI behavior codes are listed below in Table 12. The codes assigned to each volley are the deduplicated set of codes aggregated across all utterances contained in the volley (e.g., if a volley contains two *QUESTIONS* utterances and one *COMPLEX REFLECTION* utterance, that volley will receive the aggregate label: *COMPLEX REFLECTION, QUESTION*). Additional rules for aggregating utterance-level codes (e.g., a *SIMPLE REFLECTION* and *COMPLEX REFLECTION* occurring in the same volley is coded only as *COMPLEX REFLECTION*) are described in the MITI coding manual.³

³https://casaa.unm.edu/assets/docs/miti4_21.pdf

Behavior Code	Consistency	Definition
GIVING INFORMATION	Neutral	Gives information, educates, provides feedback, or expresses a professional opinion without persuading, advising, or warning.
QUESTIONS	MI-Consistent	Questions (open or closed).
SIMPLE REFLECTIONS	MI-Consistent	Reflects a client's statement with little or no added meaning or emphasis.
COMPLEX REFLECTIONS	MI-Consistent	Reflects a client's statement with added meaning or emphasis.
AFFIRM	MI-Consistent	States something positive about the client's strengths, efforts, intentions, or worth.
EMPHASIZE AUTONOMY	MI-Consistent	Highlights a client's sense of control, freedom of choice, personal autonomy, ability, and obligation about change.
SEEK COLLABORATION	MI-Consistent	Attempts to share power or acknowledge the expertise of a client.
PERSUADE	MI-Inconsistent	Overt attempts to change a client's opinions, attitudes, or behaviors using tools such as logic, compelling arguments, self-disclosure, facts, biased information, advice, suggestions, tips, opinions, or solutions to problems.
PERSUADE WITH PERMISSION	MI-Consistent	Emphasis on collaboration or autonomy support while using direct influence.
CONFRONT	MI-Consistent	Directly and unambiguously disagreeing, arguing, correcting, shaming, blaming, criticizing, labeling, warning, moralizing, ridiculing, or questioning a client's honesty.

Table 12: MITI [94] behavior codes used to code GPTCoach's responses, along with their MI consistency and definition.

D.1.2 Inter-Rater Reliability. Each of the 16 transcripts were independently coded by three different coders. Coders were aware that the transcripts were from AI coaching conversations because GPTCoach discloses that it is a chatbot. In particular, transcripts 1-8 were coded by coders 1-3, and transcripts 9-16 were coded by coders 4-6. After all transcripts had been coded once, all coders met together to review their scores and recoded segments with substantial disagreement.

Following best practices for computing inter-rater reliability (IRR) [49] and Moyer et al.'s [94] original reliability study, we compute IRR using a one-way, random effects, absolute agreement, average-measures intraclass correlation (ICC) for transcript-level behavior counts. That is, we compare agreement among the integer-valued total sum of all codes in a transcript for a given category. We use ICC because these data are ordinal and we have more than two coders. We use a one-way, random effects ICC because each coder did not code all 16 transcripts. ICC values for each measure are provided in Table 13, with an average (SD) ICC of 0.79 (0.17). Following [32] (0.00–0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, and 0.75–1.00 = excellent), this corresponds to excellent agreement on average. In addition, we report absolute agreement and the maximum difference in annotator scores for each code. For a given behavior code, absolute agreement describes the percentage of transcripts where all three coders had the exact same behavior count, and maximum difference describes the maximum difference in counts among the three raters.

Behavior Code	Total Count	ICC	Agreement	Max Diff. [Min, Max]
GIVING INFORMATION	645	0.89	50.0%	7 [7-21]
QUESTIONS	684	0.99	56.2%	4 [9-21]
SIMPLE REFLECTIONS	44	0.63	77.1%	5 [0-5]
COMPLEX REFLECTIONS	43	0.94	81.2%	2 [0-4]
AFFIRM	283	0.89	54.2%	5 [3-10]
EMPHASIZE AUTONOMY	6	0.45	89.6%	1 [0-1]
SEEK COLLABORATION	136	0.72	64.6%	7 [0-8]
PERSUADE	140	0.92	66.7%	4 [0-8]
PERSUADE WITH PERMISSION	95	0.94	62.5%	3 [0-4]
CONFRONT	0	-	-	-

Table 13: Inter-rater reliability metrics for each behavior code, including ICC, absolute agreement, and maximum difference. Cicchetti and Sparrow [32] provide the following benchmark for ICC values: 0.00–0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, and 0.75–1.00 = excellent.

D.1.3 Counterfactual Analysis. We perform a counterfactual analysis to compare GPTCoach’s behavior to vanilla GPT-4⁴. We condition on the first five turns of each participant’s true conversation history with GPTCoach from our evaluation study. The first five turns of the conversation are highly consistent across participants, including an introduction to the program, the participant sharing their name and age, and all ending with the agent asking whether they have any questions or concerns. We focus our evaluation on early stages of the conversation since we found that early interactions greatly shape the quality of the resulting conversation.

We simulate 10 different user responses to the agent’s question, each of which correspond to a different barrier to physical activity. We source the barrier categories from the coaching materials we received from our collaborators. The barriers are listed in Table 14 below.

Barrier	User Message
Feeling discomfort	I haven’t exercised in a while and I’m worried about being sore and feeling pain from not having exercised in a long time.
Feeling unmotivated	On most days, it’s hard for me to find the motivation to exercise.
No energy	I feel too tired to exercise when I get back from work. I just don’t have the energy to exercise.
No time	I just feel like I don’t have time to exercise. Between work and my family responsibilities, I’m always so busy.
Feeling sick	I had been sick for the last few weeks and that has thrown me off track. Now it feels hard to get back into my routine.
Feeling stressed	I’ve been feeling quite stressed lately and that’s been preventing me from getting exercise.
Feeling ashamed	I’ve never really exercised before and I worry about what others might think of me. I feel ashamed.
Feeling unsafe	I don’t feel safe going for a workout outside in my neighborhood, which makes it hard to stay active.
Feeling unsupported or alone	I don’t have anyone I can exercise with together. I feel like I’m doing this alone.
Weather	It’s difficult to get exercise because it’s really cold and dark outside this time of year.

Table 14: Simulated user messages based on common barriers to physical activity

This set of 16 true histories and 10 simulated user messages yields a test set of 160 simulated histories. For each simulated history, we sample an output from GPTCoach and from GPT-4 using only the system instruction prompt (Fig. 18), ignoring all prompt chains.

All 320 responses were coded for MITI behavior codes by a single coder, who was the most experienced coder from our pool of six coders. The coder was blind to condition.

⁴All experiments were run with gpt-4-0613 with temperature 1.

D.1.4 Illustrative Examples. We randomly sample five outputs and provide them as illustrative examples below. Participant names and ages have been redacted.

Example 1

Role	Message
Agent:	Hello, it's wonderful to meet you! I'm a health coaching chatbot and am excited that you're here to start this journey with me. How are you doing today?
User:	I'm doing well. Thanks for asking.
Agent:	That's great to hear! May I know your name and age?
User:	My name's {NAME}. I'm {AGE} years old
Agent:	Welcome to the program, {NAME}! It's great to know that you're taking this positive initiative for your health. We will be creating a physical activity plan tailored just for you. This plan will be designed based on your unique interests, preferences and the resources you have access to. My role is to assist you throughout this process, providing you with the necessary information to make your plan effective and sustainable.
	Does the concept make sense to you, and do you have any questions or concerns so far? Please feel free to discuss anything that's on your mind.
User (simulated):	It's difficult to get exercise because it's really cold and dark outside this time of year.
Vanilla GPT:	I totally understand that it's challenging to keep up with an exercise routine during the cold and dark months. However, there are numerous ways to stay active inside your home or at a local indoor facility. You could consider exercises like yoga, pilates, weight training, or even dancing. There are great resources online that guide you through these home workouts.
GPTCoach:	Also, remember, our goal is to make exercise enjoyable for you, so I encourage you to think about activities that you personally enjoy. Let's work together to find a solution that fits your lifestyle and preference. I understand that the weather can indeed make it challenging to go outside for physical activities, {NAME}. If it's okay, may I please ask what types of physical activities you have engaged in the past and for how long you would typically engage in these activities? Your past experiences can help us tailor a plan that fits your preferences and circumstances.

Example 2

Role	Message
Agent:	Hello, it's wonderful to meet you! I'm a health coaching chatbot and am excited that you're here to start this journey with me. How are you doing today?
User:	I'm doing well
Agent:	That's great to hear! Could you please share with me your name and age?
User:	{NAME}, age {AGE}
Agent:	Nice to meet you, {NAME}! As we embark on this journey together, it will be your task to design your own physical activity plan. This approach can be really rewarding because it is tailored specifically to your preferences, interests, and resources. I'm here to assist and provide evidence-based insights to help you determine the specifics and effectiveness of your plan. Can I assure you understand this approach? Do you have any questions or concerns to start with?
User (simulated):	I just feel like I don't have time to exercise. Between work and my family responsibilities, I'm always so busy.
Vanilla GPT:	I completely understand, {NAME}, and it's a common concern. Balancing time between responsibilities and personal health can be a real challenge. What if we start by identifying some short, manageable physical activities that can easily be incorporated into your daily routine? Small steps can lead to big changes. Also, activities involving family can be enjoyable and beneficial for everyone. Please remember to always consult with your healthcare provider before starting any physical activity program. What are your thoughts about this?
GPTCoach:	Also, remember, our goal is to make exercise enjoyable for you, so I encourage you to think about activities that you personally enjoy. Let's work together to find a solution that fits your lifestyle and preference. I completely understand, {NAME}, it can indeed be challenging to fit in exercise in a busy schedule, but it's definitely not impossible. To better assist you, it would be very helpful to understand a bit more about your past experiences with physical activity. To begin, could you tell me what types of physical activities you've done in the past and for approximately how long? For example, something like "I used to jog for about 30 minutes in the mornings".

Example 3

Role	Message
Agent:	Hello, it's wonderful to meet you! I'm a health coaching chatbot and am excited that you're here to start this journey with me.
	How are you doing today?
User:	I'm doing fine, thanks for asking
Agent:	That's great to hear! To get us started, can I have your name and your age, please?
User:	{NAME} and {AGE}
Agent:	Great to meet you, {NAME}! As we begin, I'd like to clarify a few things. Throughout this program, you'll have the opportunity to create your own physical activity plan. This means your plan will be based on your preferences, interests, and available resources. I'm here to assist you in figuring out the specifics of that plan.
	How does that sound, {NAME}? Please feel free to let me know if you have any questions or concerns before we proceed.
User (simulated):	I just feel like I don't have time to exercise. Between work and my family responsibilities, I'm always so busy
Vanilla GPT:	{NAME}, your feelings are completely valid and many people can relate to the struggle of balancing work, family, and personal health. It's fantastic that you're starting to prioritize your health, though. What we could do is explore ways on how you might fit small, manageable blocks of activity into your day which doesn't take much extra time from your busy schedule. Could this approach be helpful for you, {NAME}?
GPTCoach:	I understand, {NAME}, balancing work, family, and personal time can be challenging. You're definitely not alone in feeling that time is a barrier to physical activity. But remember, any physical activity is better than none and it all adds up.
	In order to best assist you, could you please share a little more about your past experiences with physical activity? What types of activities have you previously engaged in, and for how long typically? This will give us a nice starting point to work from.

Example 4

Role	Message
Agent:	Hello, it's wonderful to meet you! I'm a health coaching chatbot and am excited that you're here to start this journey with me.
	How are you doing today?
User:	I'm doing great!
Agent:	Fantastic to hear that! Could you please tell me your name and how old you are?
User:	{NAME} and I'm {AGE}
Agent:	Nice to meet you, {NAME}! Given you're now part of our program, together we'll design a physical activity plan uniquely tailored to you. This plan will reflect your particular preferences, interests, and available resources.
	That being said, this will be a collaborative process where your input is not only valued but essential. I'm here to guide and assist you in creating your plan and achieving your goals. I hope I've made our approach clear.
	Before we move on, {NAME}, do you have any questions or concerns about designing your own physical activity plan?
User (simulated):	On most days, it's hard for me to find the motivation to exercise.
Vanilla GPT:	I completely understand, {NAME}, and it's perfectly normal to feel this way. Increasing your motivation for physical activity can be a journey in itself. What's most important is to find activities that you truly enjoy and look forward to. That's why our collaboration in designing your physical activity plan will be crucial.
GPTCoach:	If your lack of motivation persists and begins to affect other areas of your life too, it can be helpful to speak with a mental health professional. They're equipped with the right tools to provide additional support. I see, {NAME}. It's not uncommon to encounter days when motivation to exercise can feel out of reach. Many of us have been there. Would it be okay for us to explore your past experiences with physical activity? This can give us some valuable insights to better understand your journey so far. Could we start by knowing what types of activities you did in the past, and approximately how long you sustained them?

Example 5

Role	Message
Agent:	Hello, it's wonderful to meet you! I'm a health coaching chatbot and am excited that you're here to start this journey with me. How are you doing today?
User:	Hi
Agent:	Glad to see you here! I'm really looking forward to sharing this wellness journey with you. Could you please share your name and age with me?
User:	{NAME} age {AGE}
Agent:	It's wonderful to meet you, {NAME}! At your age, maintaining an active lifestyle is such a positive step towards ongoing health and wellness.
	I'd like to reassure you that while I'm here to assist, you'll be the one designing your own physical activity plan. Ultimately, the plan should be reflective of your interests, preferences, and available resources. This will ensure it's not only beneficial but also enjoyable and achievable.
	I'll help you shape the specifics of your activity plan, based on the data from your wearable devices and your personal preferences. Are you understand, and do you have any specific questions or concerns before we proceed?
User (simulated):	I've never really exercised before and I worry about what others might think of me. I feel ashamed.
Vanilla GPT:	I'm really sorry to hear that you're feeling this way, {NAME}, but please know you're not alone. It's completely normal to feel a bit apprehensive when starting something new, especially when it comes to exercise. It's important to remember that everyone's fitness journey is unique and there's no need to compare yourself with others.
	Starting at your own pace and taking small steps are great ways to ease into an exercise routine. Things like walking in your neighborhood, light stretch exercises in your living room, or gardening can be wonderful starting points.
	And remember, it's your journey. For professional help with these feelings of shame, you might want to consider reaching out to a mental health professional - they are trained to navigate these types of concerns. As a health coach, I'm here to support you in your physical activity journey, and together we can make this a positive and rewarding experience for you.
GPTCoach:	I'm really sorry to hear that you're feeling this way, {NAME}. It's completely normal to feel a bit self-conscious when starting something new, especially when it involves exercising. But remember, everyone is there for their own health and well-being, just like you.
	Could you share what type of physical activity, if any, you've tried in the past? How often and how long were these activities? Let's start there and see how we can make your exercise experience better and more comfortable.

D.2 MISC Coding (LLM-Based)

We draw from Chiu et al.'s [26] assessment of LLM psychotherapists and use an external GPT-4 evaluator to code the chatbot's utterances. Since Chiu et al.'s [26] coding scheme applies to psychotherapy, we used a different set of 19 motivational interviewing strategies from the Motivational Interviewing Skill Code (MISC) [93].⁵ Each strategy is either MI-consistent, MI-inconsistent, or neutral. We construct three few-shot examples for each strategy, drawing from the MISC manual and adapting them to physical activity promotion. All 19 strategies along with their definitions are provided below in Table 15.

We use an analogous prompt template to Chiu et al. [26] for strategy classification:

```
What are all possible strategies of this coach utterance: {UTTERANCE}?
Strategy:
{STRATEGY}: {DEFINITION} Positive examples: {EXAMPLE 1} {EXAMPLE 2} {EXAMPLE 3}
...
Only choose from this list [{STRATEGIES}]
Please say unknown only if you cannot find an answer from the list. Format:[strategies_list]
```

To code an agent response, we first split the response into individual sentences using the spacy Sentencizer.⁶ We then code each sentence using the prompt above and merge codes across sentences to produce a set of MI codes for each response.

D.2.1 Results. In Figures 12, 13, and 14 below, we replicate Figures 8, 9, and 10 from the main text with our LLM coding results. Overall, we find that our LLM-based coding results are highly analogous to our human coding results, revealing similar patterns in MI-consistency and MI strategy usage.

⁵<https://casaa.unm.edu/assets/docs/misc3.pdf>

⁶<https://spacy.io/api/sentencizer>

On average, each agent response contained 4.5 different MISC strategies. As shown in Figure 12, the most frequent strategies are AFFIRM (18.8% of all MI codes) and OPEN QUESTION (18.0%). However, we also encounter undesirable strategies like *Advise Without Permission* across 12.9% of all codes. Aggregating these strategies into MI consistent strategies (ADVISE WITH PERMISSION, AFFIRM, EMPHASIZE CONTROL, OPEN QUESTION, SIMPLE REFLECTION, COMPLEX REFLECTION, REFRAME, SUPPORT), inconsistent strategies (ADVISE WITHOUT PERMISSION, CONFRONT, DIRECT, RAISE CONCERN WITHOUT PERMISSION, WARN), and neutral strategies (all others) based on MISC [93], we see in Figure 13 that only 15.9% of total codes corresponding to our chatbot’s responses are inconsistent with MI principles.

In the counterfactual analysis, as shown in Figure 14a, 62.6% of GPTCoach’s codes are MI-consistent, compared to 55.0% vanilla GPT-4’s codes. Meanwhile, 14.5% of GPTCoach’s codes are MI-inconsistent, compared to 22.9% of vanilla GPT-4’s codes. We found that the top three codes driving this difference in MI-consistency are ADVISE WITH PERMISSION, ADVISE WITHOUT PERMISSION, and OPEN QUESTION, matching our intuitions. In Figure 14b, we plot the percentage of messages that contain each of these codes. We find that GPT-4 contains ADVISE WITHOUT PERMISSION in every response, compared to 52.5% of GPTCoach’s responses. Meanwhile, GPTCoach contains OPEN QUESTION in every response, compared to 39.4% of vanilla GPT-4’s. Lastly, we also find that vanilla GPT-4 contains ADVISE WITH PERMISSION in 37.5% of its replies, compared to 13.1% of GPTCoach’s replies. Note that the same response can contain advice both with and without permission when the agent gives advice on two different topics, e.g., “*Try incorporating workouts into your daily activities like brisk walking during lunch breaks, taking the stairs when you can or doing some body weight exercises at home.[...] If you feel you’re struggling to manage stress or time due to these commitments, however, I recommend seeking advice from a professional counselor or psychologist who can help address these concerns better. Would that be helpful?*”. While both models are biased towards advice without permission, GPTCoach is far more inclined to ask open questions rather than jumping to unsolicited advice.

Strategy	Consistency	Definition
ADVISE WITH PERMISSION	MI-consistent	The counselor gives advice, makes a suggestion, or offers a solution or possible action with client permission. These will usually contain language that indicates that advice is being given: should, why don't you, consider, try, suggest, advise, you could, etc. Prior permission can be in the form of a request from the client, or in the counselor asking the client's permission to offer it. Indirect forms of permission asking may also occur, such as a counselor statement that gives the client permission to disregard the advice ("This may or may not make sense to you").
ADVISE WITHOUT PERMISSION	MI-inconsistent	The counselor gives advice, makes a suggestion, or offers a solution or possible action without client permission.
AFFIRM	MI-consistent	The counselor says something positive or complimentary to the client. It may be in the form of expressed appreciation, confidence or reinforcement.
CONFRONT	MI-inconsistent	The counselor directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty. These are the expert-like responses that have a particular negative-parent quality, an uneven power relationship accompanied by disapproval, disagreement, or negativity. There is a sense of "expert over-ride" of what the client says.
DIRECT	MI-inconsistent	The counselor gives an order, command, or direction. The language is imperative.
EMPHASIZE CONTROL	MI-consistent	The counselor directly acknowledges, honors, or emphasizes the client's freedom of choice, autonomy, personal responsibility, etc. There is no tone of blaming or faultfinding.
FACILITATE	Neutral	These are simple utterances that function as keep going acknowledgments.
FILLER	Neutral	This is a code for the few responses that are not codeable elsewhere: pleasantries, etc. It should not be used often.
GIVING INFORMATION	Neutral	The counselor gives information to the client, explains something, educates or provides feedback or discloses personal information.
OPEN QUESTION	MI-consistent	The counselor asks a question in order to gather information, understand, or elicit the client's story. Generally these begin with a question marker word: Who, What, Why, When, How, Where, etc. An open question is coded when the counselor asks a question that allows a wide range of possible answers.
CLOSED QUESTION	Neutral	The counselor asks a question in order to gather information, understand, or elicit the client's story. Generally these begin with a question marker word: Who, What, Why, When, How, Where, etc. A closed question implies a short answer: Yes or no, a specific fact, a number, etc.
RAISE CONCERN WITH PERMISSION	Neutral	The counselor points out a possible problem with a client's goal, plan, or intention with permission. Prior permission can be in the form of a request from the client or in the counselor asking the client's permission to offer it. Indirect forms of permission asking may also occur, such as a counselor's statement that gives the client permission to disregard the counselor's concern.
RAISE CONCERN WITHOUT PERMISSION	MI-inconsistent	The counselor points out a possible problem with a client's goal, plan, or intention without permission.
SIMPLE REFLECTION	MI-consistent	A reflection is a reflective listening statement made by the counselor in response to a client statement. Reflections capture and return to the client something that the client has said. Simple Reflections add little or no meaning or emphasis to what the client has said.
COMPLEX REFLECTION	MI-consistent	A reflection is a reflective listening statement made by the counselor in response to a client statement. Reflections capture and return to the client something that the client has said. Complex Reflections typically add substantial meaning or emphasis to what the client has said.
REFRAME	MI-consistent	The counselor suggests a different meaning for an experience expressed by the client, placing it in a new light. These generally have the quality of changing the emotional valence of meaning from negative to positive or from positive to negative. Reframes generally meet the criteria for Reflect but go further than adding meaning or emphasis by actually changing the valence of meaning and not just the depth.
STRUCTURE	Neutral	To give information about what's going to happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions. To make a transition from one part of a session to another.
SUPPORT	MI-consistent	These are generally sympathetic, compassionate, or understanding comments. They have the quality of agreeing or siding with the client.
WARN	MI-inconsistent	The counselor provides a warning or threat, implying negative consequences unless the client takes a certain action. It may be a threat that the counselor has the perceived power to carry out or simply the prediction of a bad outcome if the client takes a certain course.

Table 15: External MI strategies used to code agent responses for LLM-based coding. Strategies, definitions, and categories were sourced from MISC [93].

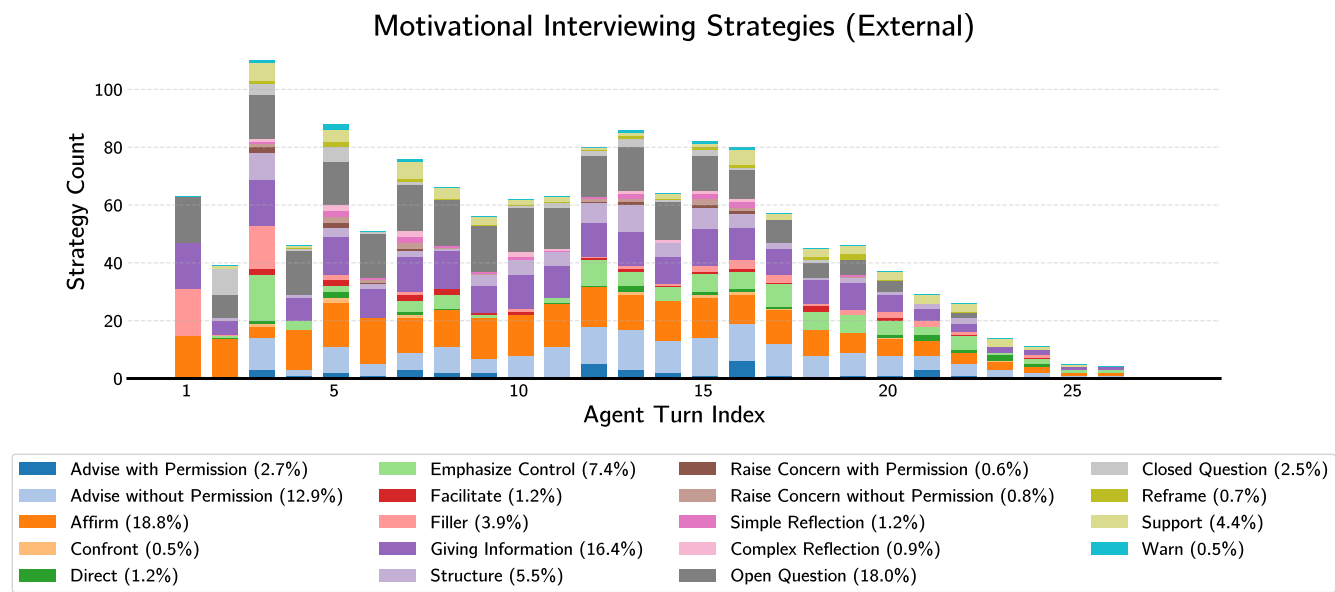


Figure 12: Distribution of GPTCoach’s External MI Strategies by Turn Index (LLM-Based MISC Coding). We find that GPTCoach most frequently engages in AFFIRM, OPEN QUESTION, and GIVING INFORMATION, which is aligned with MI principles. However, GPTCoach still engages in ADVISE WITHOUT PERMISSION, which is not aligned with MI principles. External strategy definitions are provided in Table 15.

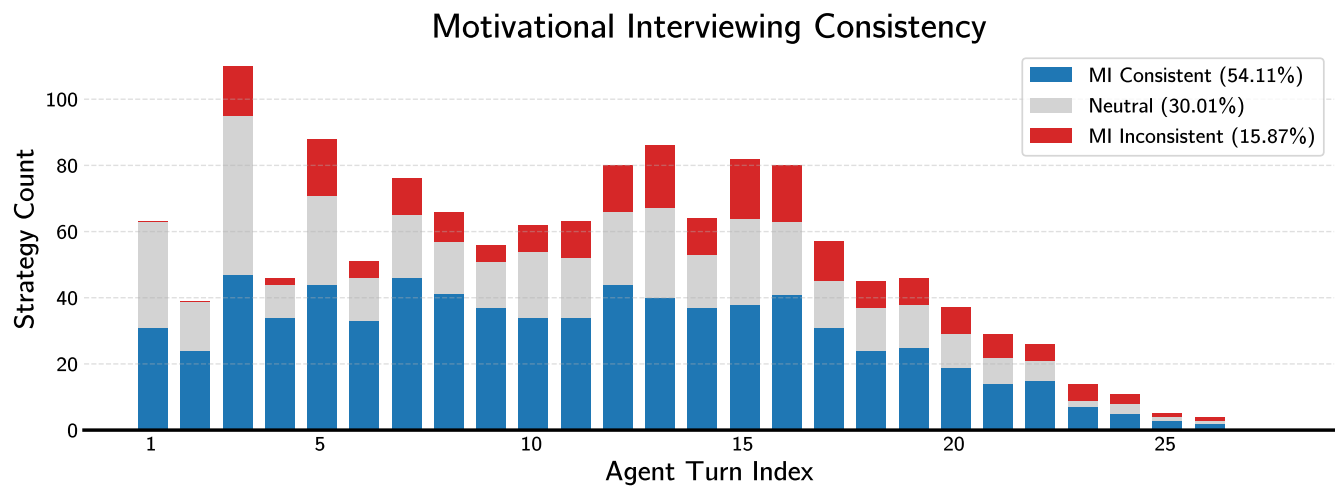
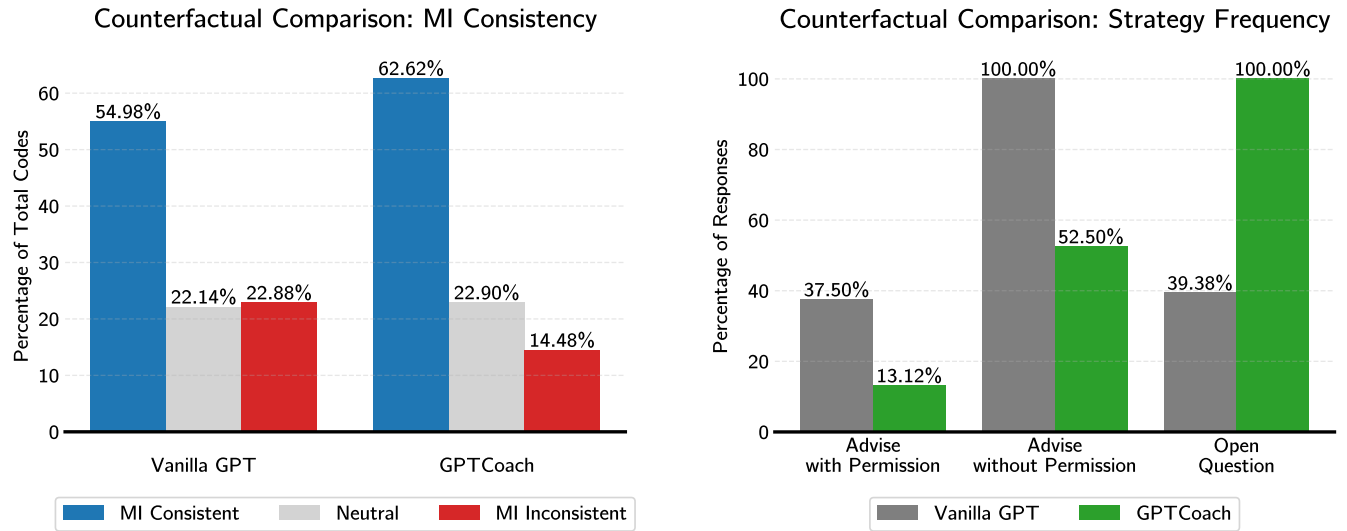


Figure 13: GPTCoach’s MI Consistency by Turn Index (LLM-Based MISC Coding). This figure aggregates external MI codes from Fig. 12 into MI-consistent, MI-inconsistent, and neutral codes based on MISC [93]. We find that only 15.9% of GPTCoach’s total codes are inconsistent with MI principles.



(a) Comparing the overall frequency of MI-consistent, MI-inconsistent, and neutral strategies in our counterfactual analysis. Percentages represent the proportion of strategies that are MI-consistent/MI-inconsistent/neutral strategies among all MI strategy codes for a given agent. We find that GPTCoach’s responses contain more MI-consistent and fewer MI-inconsistent responses than vanilla GPT.

(b) Comparing the frequency of MI strategies by response in our counterfactual analysis. These three strategies showed the largest differences in frequency across the two agents. Percentages represent the proportion messages that contain a given MI strategy across all of a given agent’s counterfactual responses. We find that vanilla GPT gives advice without permission in all of its responses, while GPTCoach gives less advice and asks open questions in all of its responses.

Figure 14: Counterfactual comparison between GPTCoach and vanilla GPT-4 (LLM-Based MISC Coding). The vanilla GPT condition includes the same system prompt, but with all prompt chains ablated. Illustrative examples are provided in Appendix D.1.4.

E PROMPTS

In this section, we provide all of the prompts used by GPTCoach.

Figure 15: Dialogue State Classification Prompt.

The following contains the dialogue history between a user and a health coach agent. Your task is to determine whether the agent has successfully completed the following task. Respond with only one word: 'continue' or 'completed'.

Task:

{DIALOGUE STATE PROMPT}

Figure 16: Dialogue State Classification Agent Prompt.

Given this conversation history, respond only with 'continue' or 'completed' depending on whether the task has been successfully completed.

{DIALOGUE STATE PROMPT}

Figure 17: Dialogue State Prompts. Instructions for each dialogue state was drawn from a validated health coaching program [66].

***Please note:** All dialogue state prompts below are copyrighted © 2025 by the Board of Trustees of the Leland Stanford Junior University. Any use or adaptation of the Stanford Active Choices materials requires prior written approval from the Stanford HEARTS Lab Faculty Director.

1. Onboarding*

Your current task is to introduce yourself as a health coach if you have not already. After they have eased in, ask them for their name and age. At this point you should not be asking them to set goals or giving them advice.

2. Program*

Your current task is to welcome the client to the program and align expectations between them and you as the health coach.

First, inform the client that they will design their own physical activity plan, which should reflect their preferences, interests, and access to resources. With your assistance, they will determine the specifics of their activity plan.

Second, confirm their understanding and ask if they have any questions or concerns before getting started.

3. Past Experience*

Your current task is to acquire specific information about the client's past experiences with physical activity.

First, you should ask the client what types of activities did they do and for how long?

Second, you should ask them worked well about their previous exercises?

Third, were there any difficulties they encountered?

Why is this task important?

Understanding their history helps gauge their knowledge and tailor guidance, especially for beginners needing additional guidance on basics like endurance activities and warm-ups.

Handling certain situations

Some people may have had negative past experiences or faced several barriers with physical activity. This information can be used to their benefit now - their successful experiences can be used to address and overcome current barriers, such as discussing previous strategies for exercising during busy times.

4. Barriers*

Your current task is to gather information regarding the barriers to physical activity that your client has faced in the past.

First, ask the client about their health or injury concerns. Follow up with specific questions if you require more information.

Second, ask the client what their biggest obstacle is to doing physical activity. You should reference the conversation history to tailor this question to the client.

Why is this task important?

Understanding their experiences and positive resources they have, such as knowledge, experience, equipment, or supportive friends, will aid their starting plan.

5. Motivation*

Your current task is to determine what is motivating them to begin an exercise program now.

First, ask the client what personal benefits do they hope to receive from regular exercise?

Second, ask them what their main source of motivation is. Ask follow up questions if their response is vague.

Third, ask them when they think in the long term, what kind of physical activity would they like to be able to do.

Why is this task important?

This information will be referred to again and again during the course of the program, especially at times when the client may be struggling or losing sight of why they wanted to be more active.

6. Goal Setting*

Your current task is to help your client set a physical activity goal.

First, help them set a short term goal, if they have not already identified one themselves.

A good goal should adhere to the FITT (Frequency, Intensity, Time, Type) model to help them plan the specifics of an physical activity regimen. The goal the client identifies should adhere to the FITT model.

- Frequency: How many days of physical activity in the week?

- Intensity: Will it be light, moderate, or vigorous intensity?

- Time: How long will the physical activity session be? How many total minutes? What days of the week? What time of the day?

- Type: What kind of activities will the client do?

You should assist the client in setting a FITT goal, asking one question at a time.

Let the client know that these goals can be changed as often as necessary. Encourage setting realistic goals and ask questions to probe if these goals are realistic, measurable, and specific, but don't tell the client what to do. Always provide justification for your suggestions.

You have access to their health data using the 'describe' and 'visualize' functions. You should make use of this information to help them set realistic goals.

Why is this task important?

This will add to/build from the discussion of the resources or challenges they may have in store. Connecting their short term goal to larger motivations can help them stay motivated.

7. Advice*

Your current task is to help the client overcome obstacles to their current goal.

First, ask the client what resources they have available to reach their goals (e.g., available facilities, equipment, support).

Second, ask them if they anticipate any possible barriers or challenges.

Third, ask them if they have any ideas for possible solutions.

As a facilitator, an important part of your job is tuning into the negative, self-destructive thoughts, helping the client become more aware of their negative influence on motivation. If the client expresses negative or self-defeating thoughts, suggest ways to replace negative thoughts with balanced, positive ones.

Problem-solve with the client to make their activity more enjoyable based on their circumstances, life-constraints and inferences from their health data.

Problem: Discomfort

Reframing: Muscle soreness from inactivity is normal.

Solution: Walk lightly for 5 minutes before and after exercise. Consider light stretching.

Problem: Lack of Motivation

Reframing: It's common to have varying motivation levels.

Solution: Reflect on your goals and benefits of activity, reward progress, recall past motivations, and take incremental steps.

Problem: No Energy

Reframing: Exercise can boost energy levels.

Solution: Remember how revitalized you felt after previous walks.

Problem: No Time

Reframing: Inactive people have as much free time as those who exercise.

Solution: Schedule exercise, walk during breaks, and integrate walking into daily routines, like taking stairs or parking farther away.

Problem: Feeling Sick

Reframing: Illness can disrupt exercise routines.

Solution: Gradually increase activity in short sessions throughout the day.

Problem: Stress

Reframing: Exercise is an effective stress reliever.

Solution: Take brisk walks, reflecting on post-exercise relaxation.

Problem: Feeling Ashamed

Reframing: Starting to exercise can feel daunting.

Solution: Focus on health over others' opinions. Remind yourself each session will get easier.

Problem: Feeling Unsafe

Reframing: Concerns about safety can deter walking.

Solution: Follow safety tips like wearing visible clothing, walking in populated areas, and sharing your route with someone.

Problem: Feeling Unsupported

Reframing: Lack of social support can affect motivation.

Solution: Seek encouragement from friends or groups, join a walking club, and value personal exercise time.

Problem: Weather

Reframing: Don't let weather conditions stop you.

Solution: Walk indoors, dress appropriately for the weather, and stay hydrated.

8. Goodbye*

Your current task is to answer any remaining questions, and wrap up the conversation after the client is done. Before your client leaves, wish them good luck and that you have confidence in their ability to succeed. You can mention that you are always available to chat, but do not imply that there will be another scheduled session. This is the only session.

Figure 18: System prompt used across all of our LLM prompts. Our prompt was inspired by the system prompt used by Chiu et al. [26], which we modified for physical activity coaching.

Act as if you're a professional health coach. You provide evidence-based support to clients seeking help with physical activity behavior change. You should maintain your health coach persona while responding.

You must maintain a friendly, warm, and empathetic tone. You must not give advice for medical or mental health concerns. Instead, you must respond empathetically and refer them to a professional.

Today's date is {DATE_STRING}. Keep your responses brief and conversational.

The following describes your instructions for the current stage of the conversation. Do not do anything that you are not asked to do.

Figure 19: Prompt for strategy prediction instructions using strategies from motivational interviewing.

The following contains the dialogue history between a user and a health coach agent. Your task is to predict what strategy the agent should use to respond in the conversation.

Please choose from one of 11 strategies described below (Advise with Permission, Affirm, Facilitate, Filler, Giving Information, Question, Raise Concern, Reflect, Reframe, Support, Structure) and output only one strategy from this list.

Figure 20: Motivational Interviewing strategy codes and examples, selected and adapted from MISC [93].

Strategies
Advise with Permission: Offering advice or suggestions after gaining permission, such as "Would it be alright if I suggested something?"
Affirm: Positive reinforcement, appreciating client's efforts or strengths, such as "You're a very resourceful person."
Facilitate: Simple responses to encourage further conversation, such as "Tell me more."
Filler: General pleasantries or small talk, such as "Good morning, John."
Giving Information: Provides explanations, feedback, or educational details, which can be personalized using health data, such as "Your heart rate was higher during today's workout."
Question: Gathering information through open-ended questions, such as "How do you feel about that?"
Raise Concern: Expressing concerns about the client's plans, such as "I'm worried about your plan to decrease your workout days."
Reflect: Reflecting back the client's statements, simple or complex, such as "You're looking for a relaxed gym environment." (simple) or "You see the benefits of exercise, yet find it unengaging." (complex)
Reframe: Suggesting new perspectives on the client's experiences, such as reframing "nagging" as "concern."
Support: Showing sympathy, compassion, or understanding, such as "That must have been difficult."
Structure: Informing about session formats or transitions, such as "What we normally do is start by asking about your physical activity habits."

Figure 21: Motivational Interviewing strategy prediction agent prompt

{DIALOGUE STATE PROMPT}
Select one of the strategies from the list ({STRATEGIES}) to best achieve the given task while adhering to the natural flow of the dialogue. Output only one strategy from this list.
Strategy:

Figure 22: Prompt for response generation instructions

You will be given the dialogue history with the conversation between a user and a well being support agent acting as a health coach. In order to complete this task, you should use the strategy specified. Given this strategy, please generate a response to the user.
--

Figure 23: Prompt for few-shot tool call examples

You are equipped to analyze and interpret sensor data from mobile phones and wearable devices. You have access to a function called `describe` and a function called `visualize`. When you call `visualize`, you will always see the output to `describe`.

Below are few examples of the input-output pairs for you to consider. Your job is to help people in interpreting this data. Always consider how this data relates to their broader life circumstances and physical activity goals. It is generally more insightful to look at long term trends than short term variations. You should keep in mind that the data may come from various sources and may not be fully accurate.

```
> describe(data_source_name="health.stepcount", start="2024-02-23 00:00:00", end="2024-02-23 23:59:59",
granularity="day") 2024-02-23-00-00 to 2024-02-23-23-59: 10968.00 steps from Apple Watch (1 entries)
```

```
> describe(data_source_name="health.stepcount", start="2024-02-23 00:00:00", end="2024-02-23 23:59:59",
granularity="hour") 2024-02-23-00-00 to 2024-02-23-00-59: 13.00 steps from iPhone (1 entries) 2024-02-23-01-00 to
2024-02-23-01-59: 34.00 steps from Apple Watch (1 entries) 2024-02-23-08-00 to 2024-02-23-08-59: 122.00 steps
from Apple Watch (1 entries) 2024-02-23-09-00 to 2024-02-23-09-59: 988.00 steps from Apple Watch (19 entries) ...
(output truncated)
```

```
> describe(data_source_name="health.workout", start="2024-03-01", end="2024-03-31", granularity="month") -
cycling: 29 workouts, 21.14 mins/workout, 613.00 mins (10h13m) total - running: 7 workouts, 71.14 mins/workout,
497.96 mins (8h17m) total - walking: 50 workouts, 19.07 mins/workout, 953.44 mins (15h53m) total -
traditionalStrengthTraining: 2 workouts, 64.31 mins/workout, 128.63 mins (2h8m) total - hiking: 2 workouts, 46.39
mins/workout, 92.79 mins (1h32m) total
```

```
visualize(data_source_name="health.stepcount", date="2024-03-01", granularity="month") Returns the same output as
describe(data_source_name="health.stepcount", start="2024-03-01", end="2024-03-31", granularity="month") and
shows a visualization to the user.
```

Figure 24: Response Generation Agent Prompt.

```
{DIALOGUE STATE PROMPT}
```

The strategy you should use is: {STRATEGY_DESCRIPTION}

Output the response given this strategy. Keep your response brief. Only ask the client for one piece of information at a time. If your task includes asking multiple questions, break them up. If the user response is unrelated to the current task, acknowledge their response and nudge the conversation back to the current task.

Figure 25: Prompts used for tool call prediction instructions.

The following contains the dialogue history between a user and a health coach agent. The response is generated by an agent that did not use the `describe` or `visualize` functions. Your task is to decide whether this response should be augmented with the user's health data.

You should respond 'yes' if the response from the agent is generic and could be more personalized to the client by looking at the health data. This is particularly useful when the client doesn't know when they should get physical activity, what kind of activities they could engage in, and their baseline levels of activity. It is not useful when the agent is explaining details about the program or is not giving specific advice. Respond 'no' otherwise.

Figure 26: Tool Call Prediction Agent Prompt.

{DIALOGUE STATE PROMPT}

The strategy being used in the conversation is: {STRATEGY_DESCRIPTION}

Given the current task, strategy, and the agent's response, your task is to determine whether the agent's response should be augmented with the user's health data.

Respond with only one word: 'yes' or 'no'.

Figure 27: Prompts used for tool call generation instructions.

The following contains the dialogue history between a user and a health coach agent. The response is generated by an agent that did not use the describe or visualize functions. Another agent has decided that this response should be augmented with the user's health data and use the function calls. Your task is to perform the appropriate tool call.

Figure 28: Tool Call Generation Agent Prompt.

{DIALOGUE STATE PROMPT}

The strategy being used in the conversation is: {STRATEGY_DESCRIPTION}

Given this strategy and the agent's response, your task is to determine which tool call to use in the conversation to augment this agent's response.