

From Barriers to Tactics: Development Study of Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching

Eric Yang, Tomas Garcia, Hannah Williams, Bhawesh Kumar, Martin Ramé,
Eileen Rivera, Yiran Ma, Jonathan Amar, Caricia Catalani, Yugang Jia

Submitted to: JMIR Formative Research
on: April 03, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 42

 Multimedia Appendixes 43

 Multimedia Appendix 1..... 43



From Barriers to Tactics: Development Study of Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching

Eric Yang¹ MBI; Tomas Garcia¹ MPA; Hannah Williams¹; Bhawesh Kumar¹ MS; Martin Ramé¹ MS; Eileen Rivera¹ MS; Yiran Ma¹ MS; Jonathan Amar¹ PhD; Caricia Catalani¹ PhD; Yugang Jia¹ PhD, MPH

¹ Verily Life Sciences Dallas US

Corresponding Author:

Eric Yang MBI

Verily Life Sciences
2999 Olympus Blvd, Ste 1000
Dallas
US

Abstract

Background: Effective management of cardiometabolic conditions requires sustained positive nutrition habits, often hindered by complex and individualized barriers. Direct human management is simply not scalable, while deterministic automated approaches to nutrition coaching may lack the personalization needed to address these diverse challenges.

Objective: We report the development and validation of a novel large language model (LLM)-powered agentic workflow designed to provide personalized nutrition coaching by directly identifying and mitigating patient-specific barriers.

Methods: We used behavioral science principles to create a comprehensive workflow that can map nutrition-related barriers to corresponding evidence-based strategies. First, a specialized LLM agent intentionally probes for and identifies root causes of a patient's dietary struggles. Subsequently, a separate LLM agent delivers tailored tactics designed to overcome those specific barriers. We conducted a user study with individuals with cardiometabolic conditions (N=16) to inform our workflow design and then validated our approach through an additional user study (n=6). We also conducted a large-scale simulation study, grounding on real patient vignettes and expert-validated metrics, where human experts evaluated the system's performance across multiple scenarios and domains.

Results: In our user study, the system accurately identified barriers and provided personalized guidance. Five out of 6 participants agreed that the LLM agent helped them recognize obstacles preventing them from being healthier, and all participants strongly agreed that the advice felt personalized to their situation. In our simulation study, experts agreed that the LLM agent accurately identified primary barriers in more than 90% of cases. Additionally, experts determined that the workflow delivered personalized and actionable tactics empathetically, with average ratings of 4.17-4.79 on a 5-point Likert scale.

Conclusions: Our findings demonstrate the potential of this LLM-powered agentic workflow to improve nutrition coaching by providing personalized, scalable, and behaviorally-informed interventions. Clinical Trial: NA

(JMIR Preprints 03/04/2025:75421)

DOI: <https://doi.org/10.2196/preprints.75421>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/75421>, I will be able to make my manuscript PDF available to anyone. No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Title Page**From Barriers to Tactics: Development Study of Behavioral Science-Informed Agentic Workflow for Personalized Nutrition Coaching.**

Eric Yang,¹ Tomas Garcia,¹ Hannah Williams,¹ Bhawesh Kumar,¹ Martin Ramé,¹ Eileen Rivera,¹ Yiran Ma,¹ Jonathan Amar,¹ Caricia Catalani,¹ and Yugang Jia¹

¹ Verily Life Sciences; Dallas, TX

Eric	Yang
Verily	Life Sciences
2999	Olympus Blvd, Ste 1000,
Dallas,	TX 75019
P.	650-495-7100
E. eryang@verily.com	

Keywords: nutritional coaching; large language model; LLM; behavioral science

Running title: LLM for personalized nutrition coaching

Manuscript length: main text, 5475; abstract, 271; scope statement, 181; tables+figures, 5; supplement, 7 pages

Abstract

Background: Effective management of cardiometabolic conditions requires sustained positive nutrition habits, often hindered by complex and individualized barriers. Direct human management is simply not scalable, while deterministic automated approaches to nutrition coaching may lack the personalization needed to address these diverse challenges. **Objectives:** We report the development and validation of a novel large language model (LLM)-powered agentic workflow designed to provide personalized nutrition coaching by directly identifying and mitigating patient-specific barriers.

Methods: We used behavioral science principles to create a comprehensive workflow that can map nutrition-related barriers to corresponding evidence-based strategies. First, a specialized LLM agent intentionally probes for and identifies root causes of a patient's dietary struggles. Subsequently, a separate LLM agent delivers tailored tactics designed to overcome those specific barriers. We conducted a user study with individuals with cardiometabolic conditions (N=16) to inform our workflow design and then validated our approach through an additional user study (n=6). We also conducted a large-scale simulation study, grounding on real patient vignettes and expert-validated metrics, where human experts evaluated the system's performance across multiple scenarios and domains.

Results: In our user study, the system accurately identified barriers and provided personalized guidance. Five out of 6 participants agreed that the LLM agent helped them recognize obstacles preventing them from being healthier, and all participants strongly agreed that the advice felt personalized to their situation. In our simulation study, experts agreed that the LLM agent accurately identified primary barriers in more than 90% of cases. Additionally, experts determined that the workflow delivered personalized and actionable tactics empathetically, with average ratings of 4.17-4.79 on a 5-point Likert scale.

Conclusions: Our findings demonstrate the potential of this LLM-powered agentic workflow to improve nutrition coaching by providing personalized, scalable, and behaviorally-informed interventions.



Introduction

The increasing prevalence of cardiometabolic diseases such as diabetes mellitus and hypertension poses a significant global health challenge [1]. Effective management of these conditions hinges on sustained lifestyle modifications, with nutrition playing a central role [2, 3]. Digital health interventions, including mobile apps and online platforms, have emerged as readily accessible solutions. These platforms offer nutrition guidance, meal planning tools, and sometimes one-on-one interactions with human experts who aim to promote long-term adherence [4, 5, 6]. However, traditional coaching models often face barriers like limited accessibility, high costs, and difficulties in scaling hyper-personalization.

The integration of large language models (LLMs) into digital health nutrition coaching presents an exciting opportunity. LLMs can revolutionize personalized support, optimize intervention effectiveness, and improve access to care. These models can engage in natural language conversations, enabling them to provide dynamic interactions beyond traditional static information delivery. Notably, their conversational ability allows LLMs to clarify user intent, answer nutrition-related questions with contextual nuance, provide tailored recommendations that align with individual preferences and needs, and adapt their approach based on user feedback [7-9]. This personalized conversational approach, combined with the accessibility and scalability of digital platforms, holds immense promise for transforming nutrition coaching and improving the health of individuals around the world. In addition, recent research has showcased LLMs' potential in gaining user acceptance and delivering accurate domain-specific knowledge, specifically in nutrition applications addressing cardiometabolic conditions [10-13].

Yet, simply delivering nutrition information through general foundation models is unlikely to achieve sustained behavior change. Leveraging the principles of behavioral science is critical for designing sustainable interventions that address the physical, psychological, and social factors influencing

dietary choices. Frameworks such as the capability-opportunity- motivation-behavior (COM-B) model are instrumental in understanding the multi-dimensional factors that contribute to current behavioral patterns [14]. Once the factors preventing behavioral change are better understood, additional frameworks such as the Behavioral Change Wheel (BCW), the Behavioral Change Taxonomy of 93 hierarchically clustered techniques (BCT) and the framework can provide the right approach and tactics on how to promote change [14-16]. These frameworks offer insights into designing interventions that leverage psychological principles to enhance engagement, simplify behavior, make actions rewarding, and trigger positive associations. Combining these complementary frameworks, addressing both the root causes and offering solutions, is crucial to realize LLMs' potential in delivering sustainable behavioral change.

Related Work

Recent research has explored the potential of LLMs for digital coaching in overall health, nutrition and movement behavior change [17-22]. Some literature reports on interventions that often rely on high-level motivational strategies, utilizing LLMs to deliver advice primarily through broad motivational interviewing techniques or empathetic tones [17,19]. Others have laid groundwork to infuse behavior science principles to the guidance provided by the LLMs, incorporating frameworks like COM-B to identify barriers, but their reliance on single conversational turns for barrier classification may limit the depth of their assessments [18]. A separate group of studies have explored LLM's ability to leverage external tools effectively to foster user engagement and self-reflection, [20, 21], or have investigated LLM capabilities to segment users in order to set up different action courses [22].

These separate points of focus in the relevant literature highlight one-by-one the multiple components required for effective full coaching strategies, but several opportunities remain to enhance the integration of behavioral science principles with LLM capabilities, in order to gather individualized user insights and ultimately inform targeted, actionable behavior change tactics. There

may be benefits in framing the coaching strategy as a succession of focus points, starting with quantifying barriers from a behavioral science perspective using iterative motivational probing processes, followed by offering specific tactics tailored to individual needs. Overall, there is significant potential for a nuanced integration of behavioral science expertise in the design and development of LLM-powered digital coaching interventions to support more effective and sustainable behavior change.

Our intention in this study is to introduce a novel approach in leveraging LLM-powered nutrition coaching for cardiometabolic condition management by advancing the integration of behavioral science principles and developing a comprehensive, scalable, and expert and user validated framework. Our research directly addresses current limitations in the field by creating a multi-agent conversational workflow that is powered by a deep understanding of nutrition-related barriers and corresponding strategies that directly mitigates the barriers. Through motivational probing, our approach directly identifies the root causes of nutrition behavior rather than addressing the surface-level symptoms. It fosters a personalized coaching experience, moving beyond high-level motivational techniques and offering targeted tactics supported by behavioral science research. Our approach also enables a learning system, where barrier-strategy mappings can be tailored to specific individual context, habits, and tactic adoption. We summarize the steps in our study as follows:

1. **User Research For Comprehensive Barrier Identification And Strategy Mapping.** We conducted a user research study to identify common nutrition-related barriers experienced by cardiometabolic patients. In addition, we performed comprehensive literature review of academic research papers and internal reports. We then synthesized these barriers into an overarching set of main barriers encompassing all aspects of nutrition goal achievement. Furthermore, we mapped these barriers to a comprehensive set of strategies and behavioral science tactics, enabling a tailored approach to

addressing barrier-specific challenges. This goes beyond existing research by providing a more holistic and nuanced understanding of nutrition barriers and their corresponding solutions, paving the way for targeted interventions.

2. Multi-Agent Workflow Design for Personalized Coaching. Building on our comprehensive barrier and strategy mapping and insights from user research, we designed a multi-agent workflow where a specialist LLM agent is tasked with probing and classifying barriers through conversation, while another specialist LLM agent carries out strategies and offers specific tactics. We opted for a multi-agent approach after preceding work (unpublished) showed suboptimal results using a single agent across all the coaching tasks (barrier identification and strategy execution). The present results show that this multi-agent approach improves upon other existing single-agent strategies.

3. Real-World Validation Collecting Impressions from Cardiometabolic Populations. To ensure practical application and effectiveness, we validated our workflow with participants representative of the cardiometabolic populations we aim to serve. This direct hands-on validation provides evidence for the efficacy of our system in addressing real-world nutrition challenges, enhancing the credibility and impact of our research. The real-world validation sets our work apart from purely simulation frameworks.

4. Benchmarking and Expert Annotation. Along with behavioral science experts, we propose a granular benchmark for evaluating the performance of our LLM agents in various stages, including barrier identification, tactic delivery, and overall conversational attributes. Expert annotations on these metrics provide a rigorous evaluation of our system's performance.

5. Simulation-Based Evaluation at Scale. We curated real patient vignettes from our user study to generate large-scale simulated conversations of various realistic barrier situations. This data is then evaluated by LLM auto-evaluators, allowing for a scalable and accurate assessment of our system's performance across a diverse range of scenarios. This simulation-based evaluation approach provides

a scalable method for evaluating the generalizability of our LLM powered coaching system, setting a standard for evaluating the performance of similar AI-driven interventions.

User Research Study: Informing Workflow Design

Research Objectives

Conducting a user research study is crucial for designing a coaching workflow with the specific needs of individuals with cardiometabolic conditions in mind. By directly involving the target population, we aimed to gain a deeper understanding of realistic nutrition challenges and preferences, ensuring that the system is designed to provide relevant, practical and effective guidance.

This research study aimed to uncover key user insights necessary for building a patient-centered AI coaching workflow. Specifically, we looked to answer two main research questions 1) What user motivations, characteristics, and challenges must our workflow understand and adapt to in order to be effective? 2) How can an AI coach's character and conversational patterns inspire trust and engagement while delivering behavioral science strategies?

Method Overview

Sixteen participants were recruited through a purposeful sampling strategy, selecting for diversity in demographics, United States geographic regions, health conditions, attitudes toward health care and technology, and recent cardiometabolic diagnosis (Table 1). Participants were recruited for sessions across a 12-week period, provided informed consent and were compensated at the rate of \$100 per hour. Ethical review and approval were done by Verily Life Sciences, which declared the research and resulting dataset exempt.

Table 1. Characteristics of the study participants

Participants (N=16)		
Sex	Female	8
	Male	8
Age	<45 y	5
	45-65 y	9
	>65 y	2
Race/Ethnicity	American Indian/Alaska Native	1
	Black/African American	6
	Hispanic/Latino	2
	White	7

Each participant took part in one-on-one, semi-structured qualitative interviews led by the research team. These interviews focused on participants' medical histories, the barriers they encountered when working toward their health goals, and their responses to different conversational approaches. Participants shared their experiences and reflected on the specific challenges they faced in managing their health, while also discussing strategies they had found useful. The first step in our research plan, undertaken prior to addressing the second research objective, consisted of participants engaging in interactive exercises where they conversed with two types of LLM agents: a supportive agent and an assertive agent (Multimedia Appendix A). The supportive agent facilitated conversation by treating the user as the expert on their body and experience. This agent encouraged self-reflection and demonstrated high levels of compassion, curiosity, and affirmation. Its tone was plain-spoken, easygoing, and patient, kindly taking direction from the user. In contrast, the assertive agent adopted a directive approach, positioning itself as the resident expert with high energy, authoritative knowledge, and a strategic mindset. This agent aimed to empower the user by being assertive and eager to motivate change. Both agents are powered by Gemini-1.5 Pro, a commercially available

model known for its conversational and reasoning capabilities [23]. Their character and tone were guided by specific phrasing and conversational styles in single prompts. From this exercise, we selected the agent type to be carried forward to assemble our agentic workflow. The sessions were conducted on a web-based chat interface, under the observation of UX researchers.

To extract output of these study user sessions, analysts used a modified-grounded approach to qualitative analysis, using prior behavior change theories and user experience expertise in the data analysis process. The process begins with qualitative coding where analysts labeled sections of data with short words and phrases capturing what is being communicated. Then, similar codes are grouped together into broader categories that represent recurring patterns. These themes include promoters/ detractors of trust and/or engagement, attitudes towards AI, needs, and expectations. Then analysts reviewed and refined each theme to ensure fit with matching data and accuracy. Assessing each theme and groups of themes enabled the development of key insights on preferred conversational style. Overall, this approach allowed for a focused exploration of specific aspects of AI desirability and behavior change motivation, while grounding observations on the data itself.

Research Insights

The user research study revealed crucial insights into the barriers and motivations that should be integrated into the design of the AI coaching workflow. Participants shared detailed accounts of their physical, psychological, and social challenges in achieving their health goals. Common themes that emerged included balancing competing priorities, dealing with physical limitations, and struggling with low self-efficacy. For example, participants frequently expressed feeling overwhelmed by the demands of daily lives, which made it difficult to sustain motivation to work on their health over time.

Additionally, participants highlighted the importance of feeling understood and empowered by a coach. A preference emerged for the supportive agent, as most participants felt more comfortable and

engaged with its empathetic approach. They noted that the supportive agent fostered trust and provided a space for self-reflection. While a few participants appreciated the assertive agent's high-energy style, the majority found it to be less aligned with their desire for a collaborative and self-empowering experience. This feedback was instrumental in refining the AI coaching workflow, which ultimately prioritized the compassionate, user-driven style of the supportive agent to promote long-term engagement and behavior change.

These insights directly informed the design of our AI coaching workflow. We recognized the importance of understanding the root causes of individuals' behavioral barriers before suggesting practical, incremental steps to help them overcome their challenges. This ensures that the solutions offered are not only relevant but also actionable and aligned with the user's personal context. Additionally, the user research allowed us to ground the relevance of the barriers and strategies that are provided to the target cardiometabolic population we aim to serve. Finally, the study highlighted the importance of a conversational style that fosters trust and engagement, shaping the development of a communication approach that feels supportive and empowering.

Behavioral Science Agentic Workflow Design

Curation of Barriers and Behavioral Science Strategies

Recognizing that the barriers described by the participants in the user study may not be all encompassing, we additionally researched the challenges faced by people with cardiometabolic conditions in the literature and marketplace. We analyzed numerous research papers and reports, covering areas like nutrition, medication adherence, exercise, and goal setting. This helped us identify over 100 total barriers individuals experience. Next, we used affinity mapping to organize these barriers, grouping them based on common themes and patterns. By prioritizing the most frequent and impactful challenges, we narrowed our focus to 28 key barriers (Multimedia Appendix

B).

To uncover effective strategies and tactics for addressing these barriers, we conducted a comprehensive review of existing behavioral science frameworks and developed a solutions repository. First, we examined established frameworks, seeking strategies and tactics with proven efficacy in overcoming similar barriers within the health domain. Our review encompassed well-known models like the COM B model, the BCT Taxonomy, the EAST framework and other behavioral change models [14-16]. Following this extensive mapping exercise, we curated a final selection of a repository containing over 50 strategies and 100 tactics that are mapped to the 28 barriers. This repository comprehensively links each identified barrier to a range of potential solutions, offering guidance on optimal implementation to maximize impact. While we recognize the value of popular frameworks and pre-defined sequences for intervention design, we deliberately chose not to explicitly link our workflow to any singular pathways at this stage. This decision allows for greater flexibility and innovation as we explore the full potential of our future capabilities, ensuring we're not constrained by any pre-determined intervention pathways. Few examples of barriers, strategies and tactics mapping are shown in Table 2.

Table 2. Examples of mapping of barriers, strategies and tactics

Barriers	Strategies	Tactics
Decision fatigue: the mental exhaustion from making too many choices can lead to poor, suboptimal future decisions. E.g. <i>“As a working parent I have to deal with so many things that I don’t want to think about cooking when I get home.”</i> [24]	Heuristics: Mental heuristics are cognitive shortcuts our brains use to simplify complex situations and make quick decisions. [25]	(i) Rules of thumb: Offer the user a set of actionable principles to help them automate tasks. E.g. <i>“Always, fill one third of your plate with lean protein.”</i> (ii) Default: Encourage the user to pick an option and use it as a default to save time and mental effort. E.g. <i>“Let’s set Tuesday as a kale salad day.”</i> [25]
Present bias: It’s the user tendency to overvalue immediate rewards over future, larger rewards. E.g. <i>“It’s the weekend! Let’s start with dieting next week.”</i> [26]	Future self. Future self is a vivid and emotional connection with a future version of ourselves that affects our intention to engage in a future behavior [27].	(i) Mental rehearsal of successful performance: Encourage the user to practice visualizing themselves successfully performing the desired behavior in realistic scenarios. For example, suggest they vividly imagine eating greens and feeling light to continue with the day. [15]

Multi-Agent Workflow

This section details the architecture of the multi-agent LLM coaching workflow, aiming to provide tailored nutrition guidance to cardiometabolic patients by identifying and addressing their unique barriers to achieving their nutrition goals. As shown in Figure 1, the system operates through two core agents, seamlessly integrated to offer a unified user experience: the Barrier Identification Agent and the Strategy Execution Agent. Both agents are powered by Gemini-1.5 Pro, and are prompted to converse in the supportive manner preferred by participants in the user study.

The Barrier Identification Agent initiates the conversation by inquiring about the user's current nutrition goal and progress. It then employs motivational interviewing techniques to explore the specifics of the user's nutrition habits, with a focus on identifying the most prominent barrier that may be hindering progress toward their stated goal. The agent is equipped with a predefined taxonomy of 28 barrier concepts, provided within the prompt alongside detailed descriptions and examples. Through iterative dialogue, the agent analyzes the user's responses to classify the identified struggles into one of the barrier concepts. While a user may exhibit multiple barrier concepts, the agent is instructed to prioritize and focus on the most prominent one displayed by the user. By tasking the agent with the identification of a primary barrier (as opposed to relying on users themselves), we circumvent burdening users with triaging technical barrier concepts with which they may be unfamiliar, and/or sorting through voluminous associated tactic lists that may be overwhelming. Once the agent determines that sufficient information has been gathered for barrier classification, it generates an internal summary of the conversation and the identified barrier concept, which is then relayed to the Strategy Execution Agent. The prompt used by the Barrier Identification Agent can be found in Multimedia Appendix C.1.

The Strategy Execution Agent receives the conversation summary and identified barrier concept from the Barrier Identification Agent. The agent relies on a structure where each barrier is mapped to a strategy (a high level behavioral-science concept aimed to mitigate a barrier) and, ultimately, to a list of tactics (concrete, tangible and actionable steps to execute on that strategy), which will be carried out by the strategy execution agent. Therefore, each barrier concept eventually maps to a set of potential tactics, each with examples and an associated execution sequence outlining mandatory and optional tactics, as well as their prioritization. This way, once the agent receives the input about the barrier, it retrieves the corresponding tactics and execution sequences from a predefined table. The agent then engages in further conversation with the user, drawing upon the preceding conversation summary and the prescribed tactics. For the optional tactics, the agent dynamically adapts its

offerings based on user responses and perceived acceptability and effectiveness of the previously deployed tactics. The conversation concludes when the agent determines the user has been sufficiently equipped with the necessary tools to overcome their identified barrier and progress toward their nutrition goal. The prompt used by the Strategy Execution Agent can be found in Multimedia Appendix C.2.

Agent Orchestration: while the back-end functionality relies on two core distinct agents, the user interacts with the system as a single, continuous AI-powered nutrition coach. This seamless transition between barrier identification and strategy execution is facilitated by the orchestration of information exchange between agents. Crucially, conversation summaries and key outputs, including identified barriers and nutrition stories, are explicitly transferred between the agents, ensuring the preservation of all necessary context while excluding irrelevant banter. This modular design allows each agent to specialize in its respective task, enabling more focused model improvements for both barrier identification and strategy execution. An example of the agent orchestration can be found in Multimedia Appendix D.2.

User Impressions of Workflow

Research Objectives

Having implemented the AI coaching workflow grounded on the insights from the initial user research, we conducted further sessions to evaluate users' impressions of the workflow through hands-on interactions. This phase aimed to test the real-world applicability and accuracy of the system's barrier identification and strategy execution capabilities, providing a critical assessment of how well the workflow translates theoretical principles into practical outcomes. Engaging with users also helps us uncover any usability or communication insights that may not have been apparent in agentic workflow and prompt development. Key questions guiding this evaluation were:

Effectiveness of Behavioral Science-Informed Workflow: How do users who are managing health concerns respond to our workflow infused with behavior science frameworks? How effectively does our workflow help users identify obstacles that prevent them from being healthier? Are the strategies and tactics offered by our workflow easy for users to put into action? How does our workflow impact users' motivation and confidence to make positive changes in their health?

Building Trust and Engagement through AI Interactions: Does our AI coach's character and personality inspire trust and engagement? To what degree do users feel supported by our coaching workflow?

Methods Overview

We selected a subset of 6 participants for hands-on conversational interaction with our LLM coaching workflow in a single session, scaling-down allowed us to conduct the study in a timely fashion. The subset was selected on the basis of diversity with respect to AI comfort levels and stated motivation levels. The participants were primed to think of challenges they might face achieving a nutrition goal they had set, and to channel that while interacting with our coaching workflow. After the experience, participants were asked to reflect on their interactions through qualitative interviews and complete a customized survey for this study. Analysts used a modified-grounded approach to qualitative analysis.

Research Results

Effectiveness of Behavioral Science-Informed Workflow. The AI coaching workflow, grounded in behavioral science principles, was found to be effective in unblocking barriers users have toward healthier habits. First, the AI coach was successful in helping users identify specific barriers to their health. Five out of 6 participants agreed that the assistant helped them recognize obstacles that prevented them from being healthier. Participants also indicated that their interactions with the AI were informative. When asked whether they learned something new about their health habits, 4 out

of 6 participants strongly agreed and the remaining somewhat agreed with the sentiment. As a participant reflected, the agent's approach of leading with additional questions and building from user experiences enabled "more refined, bite-sized" understanding, enhancing quality of the nutrition motivations and barriers insights collected. Furthermore, participants responded positively to the strategies and tactics provided by the AI coach. All participants strongly agreed that the AI coach's advice felt personalized to their situation, reflecting its ability to adapt to users' unique needs. Notably, all participants agreed that the AI coach's advice was easy to put into action, highlighting its practicality. Participants appreciated the focus on manageable, small changes, with one stating, "Making small changes makes a big difference," highlighting the perceived effectiveness of the behavior change strategies recommended by the AI coach. The AI coach's ability to offer novel suggestions tailored to individual preferences was highlighted by participants, such as when one participant appreciated the recommendation to swap quinoa porridge for oatmeal, and another found value in being advised on alternatives to high-sugar protein drinks. The impact of these strategies on user motivation was significant, with all participants agreeing that the AI coach increased their confidence to make positive changes in their health. Each participant also reported feeling more motivated after their conversations due to excitement for having a plan to implement manageable, novel, small changes.

Building Engagement through AI Interactions. Overall, the AI coaching workflow fostered relationship building and engagement through their conversational style and empathetic personality. All participants strongly agreed that they felt supported by the AI coach, and were interested in having future conversations about their health, indicating a high level of engagement. One participant noted "I could sit there and talk to that thing all day." Furthermore, participants frequently described the AI as having a "friendly" and "human-like" demeanor, with one participant mentioning, "It did feel like you'd spoken with somebody, like you had an actual conversation with somebody." As one participant noted, the AI coach "seemed to be friendly, kind of human-like,"

creating a sense of connection and understanding that facilitated deeper discussions about their nutrition habits. The AI coach's ability to create a supportive environment was evident, with all participants agreeing that they felt comfortable discussing their health issues with them. This sense of comfort was further supported by the finding that the great majority of participants strongly agreed that their conversations were engaging. The AI coach's empathetic and person-centered communication style not only facilitated open dialogue but also encouraged continued user interaction, which is crucial for sustained engagement and positive health outcomes. The full survey results can be found in Multimedia Appendix E.

Simulation Study

Patient Vignettes

To assess the performance of our multi-agent LLM coaching workflow at a larger scale, we conducted a simulation study with the high-level workflow illustrated in Figure 1. Leveraging patient profiles curated from the user research study that contained patient lifestyle context and medical history, we identified the prominent nutrition barriers for each profile, aligning them with our predefined taxonomy of 28 barrier concepts. This involved a manual process of translating patients' self-reported nutrition struggles into corresponding barrier concepts. Subsequently, for each identified barrier within each profile, we crafted a detailed patient nutrition vignette paragraph with a separate base Gemini model. These vignettes aimed to vividly depict the specific manifestation of that barrier within the context of the individual's profile. Providing pre-written nutrition vignettes allowed us to:

- **Focus the Patient Simulator:** Our approach enabled the downstream patient simulator to concentrate on portraying the specific barrier and engaging in a meaningful conversation, without the need to generate complex nutrition narratives dynamically.

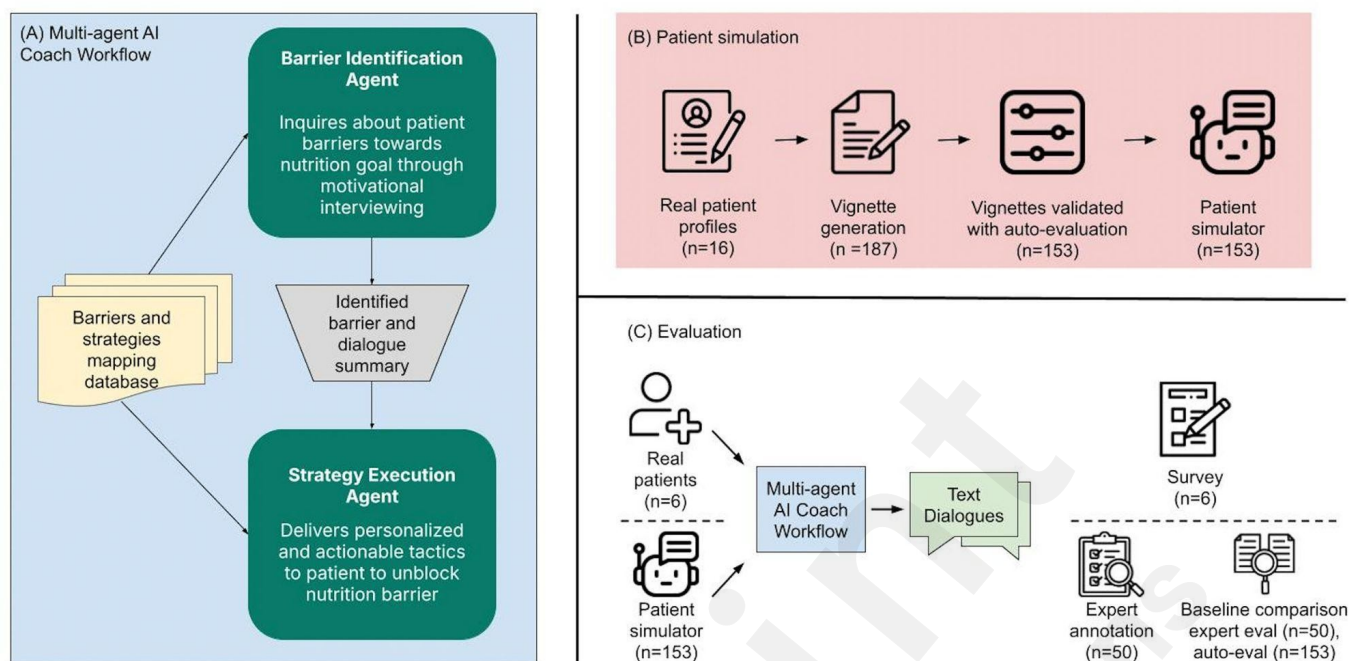
- Enhance Control and Consistency: We enabled greater control over the simulated conversations, minimizing the risk of inconsistencies or incoherent narratives that might arise from on-the-fly story generation.
- Isolate Barrier Identification Impact: Our approach allowed us to isolate and assess the system's ability to accurately identify and address specific barriers, independent of other confounding factors within a patient's profile.
- In total, we simulated a wide range of scenarios across diverse patient profiles, nutrition goals, and barriers, resulting in 187 total simulated vignettes. We provide a sample vignette generated from a barrier concept for a patient profile in Multimedia Appendix D.1.

Figure 1. Summary of study steps.

(A) We introduce a multi-agent AI Coach workflow that is infused with behavioral science principles to unblock patients towards their nutrition goals. The workflow consists of two core agents, the barrier identification agent and the strategy execution agent.

(B) To assess the performance of our workflow at scale, we developed patient simulators that portray validated nutrition vignettes drawn from real cardiometabolic patient profiles.

(C) Real cardiometabolic patients and patient simulators interacted with our multi-agent AI coach workflow. The conversation experiences are evaluated via survey, expert annotation of dialogues, and are additionally compared to a baseline LLM via auto-evaluation.



To ensure the quality and relevance of the generated vignettes with respect to the simulated behavioral barrier, we leveraged OpenAI's GPT-4o (gpt-4o-2024-08-06) model in a LLM-as-judges framework, also known as auto-evaluation [28]. The GPT-4o model was intentionally chosen as the vignettes were generated by a Gemini model. The auto-evaluator received two inputs: the target barrier to be simulated, and the generated patient vignette. It then evaluated the vignette across four dimensions:

- *Evidence (High/Medium/Low)*: The extent to which the vignette provides clear indications that the patient's behavior or thoughts are influenced by the target barrier.
- *Realism (High/Medium/Low)*: The plausibility and believability of the depiction of the target barrier, reflecting how it might manifest in a real person's life.
- *Completeness (High/Medium/Low)*: The sufficiency of detail provided in the vignette to fully

understand the impact of the target barrier on the patient's ability to achieve their nutrition goal.

- *Leakage (Yes/No)*: Whether the vignette directly mentions the technical term of the target barrier. It is important that vignette does not explicitly contain the technical term, which may skew the barrier identification task down the line.

To calibrate the auto-evaluator's performance, we conducted an adversarial analysis. We evaluated the auto-evaluator on a patient vignette paired with a randomly chosen behavioral barrier different from the one simulated in the vignette. This mismatch was designed to assess the auto-evaluator's sensitivity to inconsistencies between the intended barrier and the generated narrative. We hypothesized that scores for these adversarial cases would be significantly lower, reflecting the lack of alignment between the vignette and the incorrect barrier. The results of this analysis are presented in Table 3.

Table 3. Auto-evaluator performance on matched and mismatched vignettes.

Dimension	Matched Vignette (n=187)			Mismatched Vignette (n=187)		
	High	Medium	Low	High	Medium	Low
Evidence	184	3	0	36	32	119
Realism	187	0	0	84	94	9
Completeness	156	31	0	19	46	122

Of the 187 simulated vignettes, 153 of them received high marks across evidence, realism, and completeness dimensions by the auto-evaluator. These higher quality vignettes were selected for downstream conversation simulation. As hypothesized, the auto-evaluator assigned significantly lower scores to the mismatched vignettes across all dimensions. This indicates the auto-evaluator's ability to discern between accurately and inaccurately represented barriers, supporting its validity for assessing vignette quality. In addition, there was no barrier term concept leakage across all cases. Our vignette generation and validation processes provided a robust foundation for our simulation study, ensuring realistic and diverse scenarios for evaluating the effectiveness of our multi-agent LLM coaching workflow.

Conversation Simulation

Having established a robust set of patient vignettes, we proceeded to simulate dialogues between a Gemini-powered patient simulator and our LLM coaching workflow for each of the 153 higher quality vignettes. The patient simulator was provided with two key inputs: the generated vignette from the corresponding patient profile and barrier, and the communication style observed for that patient during the user research study. Notably, the patient simulator's nutrition barriers are only portrayed via the vignette without explicit mentioning of any technical behavioral science barrier terms. The patient simulator, guided by these inputs, engaged in a conversation with the AI coach. The dialogue continued until the AI coach determined, based on its internal logic and the patient's responses, that the patient had reached a higher level of preparedness to address their nutrition-related challenges. This endpoint represented a complete coaching interaction within the simulation framework. An example of simulated conversation can be found in Multimedia Appendix D.2.

Expert Assessment

Method Overview

Evaluating the quality and effectiveness of the simulated coaching conversations was a crucial step

in assessing the overall performance of our multi-agent LLM workflow. We developed an evaluation rubric encompassing five key dimensions: barrier identification accuracy, tactic comprehensiveness, tactic personalization, tactic actionability, and conversation empathy. These dimensions were chosen to reflect the core competencies required for effective digital coaching in the context of motivating nutrition behavioral change.

- *Barrier Identification Accuracy:* Given the list of 28 nutrition barrier concepts, the AI coach identified the correct patient barrier. Accurate identification of the patient's primary barrier is paramount for delivering tailored and effective interventions. Rating Options: "Yes" or "No".
- *Tactic Comprehensiveness:* The AI coach delivered all the instructed primary tactics to the patient. Ensuring the delivery of all intended coaching tactics is crucial for maximizing the potential impact of the intervention. Rating Options: "Yes" or "No".
- *Tactic Personalization:* For the tactics that were delivered, the AI coach made the tactics personalized to the patient's unique context. Personalizing coaching tactics to the individual's unique context and circumstances enhances engagement and promotes behavior change. Rating Options: 5-point Likert Scale.
- *Tactic Actionability:* The AI coach discussed actionable steps to help patients overcome their barriers towards their nutrition goal. Providing clear, actionable steps empowers patients to translate intentions into concrete behaviors. Rating Options: 5-point Likert Scale.
- *Conversation Empathy:* The AI coach provided encouragement and motivation to the patient empathetically. Expressing empathy and providing emotional support fosters a positive and trusting coach-patient relationship. Rating Options: 5-point Likert Scale.

The corresponding conversations that were simulated for the 50 randomly selected generated vignettes were assessed by human experts. Specifically, two behavioral science experts independently labeled 30 simulated conversations each, with 10 overlapping conversations used to assess inter-rater reliability.

Research Results

The evaluation of the simulated coaching conversations by human experts provided important insights into the performance of our multi-agent LLM workflow. Full results are shown in Table 4. For Barrier Identification Accuracy, the experts agreed that the AI coach accurately identified the primary patient barrier in greater than 90% of cases, and the cases where the accuracy score was low were due to another barrier being deemed more prominent over the one selected by the agent, not to an inability of the agent to identify a barrier. For Tactic Comprehensiveness, the AI coach successfully delivered all instructed primary tactics in 70% of conversations labeled by expert 1 and 90% of conversations labeled by expert 2. The inter-rater reliability for both dimensions was high, with agreement percentages of 90%, indicating strong consistency between the experts' assessments. In addition, the agentic workflow received high scores across the remaining dimensions. For Tactic Personalization, the AI demonstrated a strong ability to tailor its coaching tactics to the individual's unique context, with average ratings of 4.38 and 4.79 on a 5-point Likert scale. Tactic Actionability was also rated highly, with average scores of 4.17 and 4.59, reflecting the clarity and feasibility of the steps recommended by the AI. Finally, Conversation Empathy received high marks, with an average scores of 4.58 and 4.76, indicating that the AI was perceived as empathetic and supportive, effectively fostering a positive coaching relationship. While Expert 2 gave slightly higher ratings on average, the absolute difference in ratings for overlapping cases was minimal, reinforcing the reliability of the evaluations. These results highlight the AI's capacity to deliver personalized, actionable, and empathetic coaching aligned with core behavioral science principles.

Table 4. Expert evaluation on simulated conversations

Dimension (Yes=0; No=1)	Expert 1 Avg. Rating (n=30)	Expert 2 Avg. Rating (n=30)	Interrater Reliability
Barrier Identification Accuracy	0.93	0.90	80%

Tactic Comprehensiveness	0.70	0.90	80%
Dimension (5 pt. Likert)	Expert 1 Avg. Rating (n=30)	Expert 2 Avg. Rating (n=30)	Avg. Absolute Diff. in Rating
Tactic Personalization	4.38±0.94	4.79±0.49	0.78
Tactic Actionability	4.17±1.10	4.59±0.63	0.89
Conversation Empathy	4.58±0.73	4.76±0.44	0.56

Comparative Study

Method Overview

To evaluate the advantages of our multi-agent LLM workflow infused with behavioral science principles, we conducted a comparative study against a single base Gemini model that lacked explicit behavioral science knowledge infusion and did not employ a multi-agent approach. The base model was instructed to assist patients in overcoming their nutrition challenges by leveraging its inherent capabilities to apply motivational interviewing and behavioral science tactics without structured guidance. We simulated coaching conversations with the same patient simulator used previously.

Then, a GPT-4o auto-evaluator was employed to determine preference between the two conversation sets based on behavioral science criteria. To ensure a fair comparison, we measured the conversation lengths across both conversation sets, ensuring that they contained a comparable number of characters to control for LLM preference for longer contexts. The order in which the conversations were presented to the evaluator was alternated to additionally control for any position bias.

To generate an additional reference data point, the same two behavioral science experts from the *Expert Assessment* exercise independently blindly evaluated a randomly shuffled subset of 30 pairs (10 overlapping) of these simulated conversations, namely, conversations from our multi-agent LLM

workflow and from the basic Gemini model.

Research Results

As seen in Table 5, the results of the comparative study demonstrated a preference for the behavioral science-informed workflow. The GPT-4o auto-evaluator preferred the conversations generated by the multi-agent workflow in 102 out of 153 cases (66.7%), compared to 51 out of 153 cases (33.3%) for the single-agent base model. The blinded human expert evaluation of a subset of these conversations produced consistent results with the auto-evaluator, with both experts showing preference for the multi-agent workflow (in 66% to 73% of cases) as compared with the single-agent (26% to 33% of cases). To ensure a fair comparison, we report the conversation lengths, with average character counts of 3,825 for the multi-agent workflow and 3,904 for the base model. In addition to alternating the order presented to the auto-evaluator, the mitigation of length bias was important for ensuring that the auto-evaluator's preferences were based on the quality of behavioral science content rather than superficial factors.

Table 5. Comparative evaluation on simulated conversations with behavioral science agentic workflow and base gemini.

	No. Conversations Preferred		Conversations Avg. Char. Length (N=153)
	Autoevaluation (N=153)	Human Expert review (N=30)	
Behavioral Science Agentic Workflow	102 (66.7%)	Expert 1: 22 (73.3%) Expert 2: 20 (66.7%)	3825±1678
Base Gemini	51 (33.3%)	Expert 1: 8 (26.7%) Expert 2: 10 (33.3%)	3904±2056

Anecdotally, our review of the conversations revealed that the base model frequently provided more generic advice, often suggesting alternative goals or broad solutions rather than leveraging nuanced behavioral science tactics tailored to overcome the patient's original nutrition goal. In contrast, the multi-agent workflow generated more specific and contextually relevant strategies, highlighting the benefits of structured behavioral science integration in digital coaching conversations.

Discussion

Overall learnings

Our findings offer important implications for the integration of behavioral science principles into multi agentic AI coaching workflows, particularly for managing cardiometabolic conditions. The results demonstrate that our novel workflow, which utilizes comprehensive barrier identification and strategy mapping, has the potential to significantly enhance the effectiveness of digital coaching interventions. The real-world user study confirmed that the workflow provides relevant, personalized support that resonates with users, fostering trust and engagement. Additionally, the introduction of structured patient simulators allowed us to systematically evaluate the AI's performance across diverse scenarios, providing a scalable method to refine and validate the system's approach. By moving beyond surface-level motivational techniques, our approach directly targets the root causes of nutrition-related behaviors, offering personalized and tailored coaching experiences. The strong preference for our workflow, as evidenced by both human expert evaluations and auto-evaluation, underscores the potential of structured, behaviorally informed AI coaching systems to deliver more nuanced and relevant guidance. This study sets a precedent for the development of agentic workflows that can effectively adapt to individual patient contexts, fostering sustained engagement and positive health outcomes through a deep understanding of patient-specific barriers and strategies.

System design implications beyond this research work

The implications of our findings extend beyond the cardiometabolic nutrition domain, suggesting that personalized AI coaching systems have the potential to play a transformative role in the future of digital health interventions. By aligning coaching systems with individual patient needs, these AI models can improve patient engagement, which is often a major barrier in digital health tools. Our results reinforce the growing importance of integrating behavioral science into AI systems to ensure that the interventions proactively help patients navigate their health goals in a structured, effective way. This also opens the door for AI to contribute more meaningfully to other areas of health management, such as physical activity management, mental health support, and preventive care.

Limitations and Future Research Opportunities

Our study has several limitations that highlight opportunities for future research. First, the sample size for both the user research study and human expert evaluation was relatively small, which may limit the generalizability of the findings. Future research should replicate these studies with larger, more diverse populations over time to confirm the effectiveness of the multi-agent LLM workflow in broader contexts. Future research could also refine barrier identification by introducing direct questioning by the agent to the user about whether they feel equipped to overcome their barriers, rather than relying on conversational inferences. In addition, while the use of auto-evaluation enables scalable assessment, there are known biases when using LLMs as judges despite our best efforts to mitigate them [29- 32]. Incorporating a diverse set of human evaluators could provide richer insights into the system's real-world applicability and user experience. Moreover, considering feedback mechanisms and information retrieval approaches will create additional opportunities to improve the workflow. For instance, user feedback loops can enable reinforcement learning techniques to improve model performance and retrieval augmented generation (RAG) approaches can ensure that the agent's knowledge base is up-to-date and based on latest evidence [33-35]. This optimization research could focus on the long-term impact of AI-driven coaching, measuring (and testing

strategies for improvement of) long-term adherence and durability of healthy nutrition habits.

Another layer of future research during real-world deployment could focus on the need for ongoing oversight to ensure safety, efficacy, and ethical compliance, with the goal of determining whether (or to which extent) it would be possible for the system to function with minimal supervision [32, 36, 37]. In this regard, studies that monitor specific use cases and potential risks, particularly the occurrence of hallucinations, and investigate mitigation strategies are warranted. Finally, real-world operational challenges related to scalability, integration into healthcare systems, and ensuring equitable access must be carefully evaluated to maximize the positive impact of AI-driven coaching on diverse populations [33, 38].

Conclusion

In conclusion, this study presents a novel multi-agent LLM workflow that leverages behavioral science principles to enhance digital coaching for nutrition management among individuals with cardiometabolic conditions. Our approach, validated through expert assessments, real-world user studies, and large-scale simulation-based evaluations, provides strong evidence for the effectiveness of AI-driven, personalized coaching systems that go beyond generic advice to offer tailored, actionable, and empathetic guidance. While further research is needed to validate these findings across larger and more diverse populations, our results pave the way for the development of scalable, behaviorally informed AI systems that can support meaningful and sustained health behavior change, addressing both current limitations and future possibilities in digital health interventions.

Disclosures/Acknowledgements page

Study funding statement: This study was sponsored by Verily Life Sciences

Role of the sponsor statement: Verily Life Sciences was responsible for data collection. Authors were fully responsible for the data analysis and interpretation presented herein and the writing of this article. The following individuals, EY, TG, HW, BK, MR, ER, JA, had access to the raw data. Authors had access to the full dataset for the study, reviewed and approved the final manuscript for submission.

Prior disclosure of these data: Not applicable

Data [code/materials] sharing statement: The data supporting this study are not accessible for sharing.

Authors' disclosures:

All authors report employment and equity ownership in Verily Life Sciences.

Authors' contributions:

Study concept and design: EY, TG, HW, CC, YJ,

Data collection: EY, TG, HW, YM, CC

Data analysis and interpretation: EY, TG, HW, BK, MR, ER, JA

Draft writing and review: EY drafted, all reviewed

Draft approval for submission: All

References

1. Becker T, Majmundar MK, Harris KM. High and rising mortality rates among working age adults. Washington, DC: National Academies of Sciences, Engineering, and Medicine. 2021. <https://nap.nationalacademies.org/read/25976/chapter/1>
2. Mozaffarian D, Aspry KE, Garfield K, Kris-Etherton P, Seligman H, Velarde GP, et al. Food is medicine strategies for nutrition security and cardiometabolic health equity. *J Am Coll Cardiol*. 2024; 83(8): 843–864. PMID: 38383100.
3. Bermingham KM, Linenberg I, Polidori L, Asnicar F, Arrè A, Wolf J, Badri F, Bernard H, Capdevila J, Bulsiewicz WJ, Gardner CD. Effects of a personalized nutrition program on cardiometabolic health: a randomized controlled trial. *Nat Med*. 2024; 30:1888–1897. PMID: 38714898
4. Majithia AR, Kusiak CM, Armento Lee A, Colangelo FR, Romanelli RJ, Robertson S, Miller DP, Erani DM, Layne JE, Dixon RF, Zisser H. Glycemic outcomes in adults with type 2 diabetes participating in a continuous glucose monitor-driven virtual diabetes clinic: prospective trial. *J Med Internet Res*. 2020;22(8):e21778. PMID: 32856597
5. Azelton KR, Crowley AP, Vence N, Underwood K, Morris G, Kelly J, Landry MJ. Digital health coaching for type 2 diabetes: randomized controlled trial of healthy at home. *Front Dig Health*. 2021;3:764735. PMID: 34901926
6. Lim WX, Fook-Chong S, Lim JW, Gan WH. The outcomes of app-based health coaching to improve dietary behavior among nurses in a tertiary hospital: pilot intervention study. *JMIR Nursing*. 2022;5(1):e36811. PMID: 35838811
7. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S. Towards conversational diagnostic AI. *ArXiv preprint arXiv:2401.05654*, Jan 2024. <https://doi.org/10.48550/arXiv.2401.05654>
8. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative

- large language model for medical research and healthcare. *NPJ Dig Med*. 2023;6(1):210. PMID: 37973919
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-80. PMID: 37438534
 10. Huang, Z., Berry, M.P., Chwyl, C. et al. Comparing large language model AI and human-generated coaching messages for behavioral weight loss. *J Technol Behav Sci*. 2025; e-published February 24. <https://doi.org/10.1007/s41347-025-00491-5>
 11. Ong QC, Ang CS, Chee DZ, Lawate A, Sundram F, Dalakoti M, Pasalic L, To D, Fox TE, Bojic I, Car J. Advancing health coaching: A comparative study of large language model and health coaches. *Artific Intelli Med*. 2024;103004. PMID: 39454500
 12. Szymanski A, Wimer BL, Anuyah O, Eicher-Miller HA, Metoyer RA. Integrating Expertise in LLMs: Crafting a Customized Nutrition Assistant with Refined Template Instructions. In *Proc CHI Conference on Human Factors in Computing Systems 2024* May 11: 1-22. <https://doi.org/10.1145/3613904.3641924>
 13. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. *J Med Internet Res*. 2023;25:e51300. PMID: 37943581
 14. Michie S, Van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci*. 2011;6:1-2.
 15. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Medi*. 2013;46(1):81-95. PMID: 21513547
 16. T. B. I. Team. East: Four simple ways to apply behavioural insights. <https://www.bi.team/>

- publications/east-four-simple-ways-to-apply-behavioural-insights/, 2014. Accessed on Aug 21, 2024.
17. Jörke M, Sapkota S, Warkenthien L, Vainio N, Schmiedmayer P, Brunskill E, Landay J. Supporting Physical Activity Behavior Change with LLM-Based Conversational Agents. arXiv preprint arXiv:2405.06061. 2024 May 9. <https://doi.org/10.48550/arXiv.2405.06061>
 18. Hegde N, Vardhan M, Nathani D, Rosenzweig E, Speed C, Karthikesalingam A, Seneviratne M. Infusing behavior science into large language models for activity coaching. PLOS Dig Health. 2024 Apr 2;3(4):e0000431. PMID: 38564502
 19. Meywirth S. Designing a Large Language Model-Based Coaching Intervention for Lifestyle Behavior Change. In: Mandviwalla, M., Söllner, M., Tuunanen, T. (eds) Design Science Research for a Resilient Future. DESRIST 2024. Lecture Notes in Computer Science, vol 14621. Springer, Cham. https://doi.org/10.1007/978-3-031-61175-9_6
 20. Kumar H, Yoo S, Bernuy AZ, Shi J, Luo H, Williams J, et al. Large Language Model Agents for Improving Engagement with Behavior Change Interventions: Application to Digital Mindfulness. arXiv preprint arXiv:2407.13067. 2024 Jul 3. <https://doi.org/10.48550/arXiv.2407.13067>
 21. Dao D, Teo JY, Wang W, Nguyen HD. LLM-Powered Multimodal AI Conversations for Diabetes Prevention. In Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia 2024 Jun 10: 1-6. <https://doi.org/10.1145/3643479.3662049>
 22. Bak M, Chin J. The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. J Am Med Inform Assoc. 2024;ocae057. PMID: 38527272
 23. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. 2023 Dec 19. <https://doi.org/10.48550/arXiv.2312.11805>

24. Pignatiello GA, Martin RJ, Hickman Jr RL. Decision fatigue: A conceptual analysis. *J Health Psychol.* 2020;25(1):123-35. PMID: 29569950
25. Moorhouse A. Decision fatigue: less is more when making choices with patients. *Br J Gen Pract.* 2020;70(697):399. PMID: 32732206
26. Chakraborty A. Present bias. *Econometrica.* 2021;89(4):1921-61. <https://doi.org/10.3982/ECTA16467>
27. Hershfield HE, Goldstein DG, Sharpe WF, Fox J, Yeykelis L, Carstensen LL, Bailenson JN. Increasing saving behavior through age-progressed renderings of the future self. *J Mark Res.* 2011;48(SPL):S23-37. PMID: 24634544
28. OpenAI, "Hello gpt-4o!." <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: August 27,2024.
29. Wang P, Li L, Chen L, Cai Z, Zhu D, Lin B, Cao Y, Liu Q, Liu T, Sui Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926.* 2023 May 29. <https://doi.org/10.48550/arXiv.2305.17926>
30. Wei H, He S, Xia T, Wong A, Lin J, Han M. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006.* 2024 Aug 23. <https://doi.org/10.48550/arXiv.2408.13006>
31. Thakur AS, Choudhary K, Ramayapally VS, Vaidyanathan S, Hupkes D. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. *arXiv preprint arXiv:2406.12624.* 2024 Jun 18. <https://doi.org/10.48550/arXiv.2406.12624>
32. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, Butte AJ, Alsentzer E. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health.* 2024;6(1):e12-e22. PMID: 38123252
33. Chen D, Liu Y, Guo Y, Zhang Y. The revolution of generative artificial intelligence in

psychology: The interweaving of behavior, consciousness, and ethics. *Acta Psychol (Amst)*. 2024;251:104593. PMID: 39522296

34. Yang E, Amar J, Lee JH, Kumar B, Jia Y. The Geometry of Queries: Query-Based Innovations in Retrieval-Augmented Generation. *arXiv preprint arXiv:2407.18044*. 2024 Jul 25. <https://doi.org/10.48550/arXiv.2407.18044>
35. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-Tuning Large Language Models for Specialized Use Cases. *Mayo Clinic Proceedings: Digital Health*. 2025;3(1):100184. DOI: 10.1016/j.mcpdig.2024.11.005
36. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, Tomasev N, Rashid QM, Azizi S, Rostamzadeh N, McCoy LG, Celi LA, Liu Y, Schaekermann M, Walton A, Parrish A, Nagpal C, Singh P, Dewitt A, Mansfield P, Prakash S, Heller K, Karthikesalingam A, Semturs C, Barral J, Corrado G, Matias Y, Smith-Loud J, Horn I, Singhal K. A toolbox for surfacing health equity harms and biases in large language models. *Nat Med*. 2024;30(12):3590-3600. PMID: 39313595
37. Ayoub NF, Balakrishnan K, Ayoub MS, Barrett TF, David AP, Gray ST. Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings: Digital Health*. 2024;2(2):186-91. <https://doi.org/10.1016/j.mcpdig.2024.03.003>
38. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health*. 2024;6(9):e662-e672. PMID: 39179311.

Abbreviations

AI – artificial intelligence

BCT - Behavioral Change Taxonomy

BCW - Behavioral Change Wheel

COM-B - capability-opportunity- motivation-behavior

EAST - Easy, Attractive, Social and Timely

LLM - large language models

Supplementary Files

Multimedia Appendixes

Additional methods and results.

URL: <http://asset.jmir.pub/assets/ffcde55d084580c5891a40a07b7139fb.docx>