The Anatomy of a Personal Health Agent

A. Ali Heydari[1,*], Ken Gu[‡,*], Vidya Srinivas[‡,*], Hong Yu[1,*], Zhihan Zhang[‡], Yuwei Zhang[‡], Akshay Paruchuri[‡], Qian He[1], Hamid Palangi[1], Nova Hammerquist[1], Ahmed A. Metwally[1], Brent Winslow[1], Yubin Kim[‡], Kumar Ayush[1], Yuzhe Yang[1], Girish Narayanswamy[‡], Maxwell A. Xu[‡], Jake Garrison[1], Amy Armento Lee[1], Jenny Vafeiadou[1], Ben Graef[°], Isaac R. Galatzer-Levy[2], Erik Schenck[1], Andrew Barakat[1], Javier Perez[1], Jacqueline Shreibati[1], John Hernandez[1], Anthony Faranesh[1], Javier L. Prieto[‡], Conor Heneghan[1], Yun Liu[1], Jiening Zhan[‡], Mark Malhotra[1], Shwetak Patel[1], Tim Althoff[1], Xin Liu[1,†], Daniel McDuff[1,†], Xuhai "Orson" Xu[1,3,†,‡]

[1]Google Research, [2]Google DeepMind, [3]Columbia University
[*]Equal Contribution, [†]Equal Leadership
[‡]Work done while at Google Research, [°]Work done at Google via Vituity

Health is a fundamental pillar of human wellness, and the rapid advancements in large language models (LLMs) have driven the development of a new generation of health agents. However, the application of health agents to fulfill the diverse needs of individuals in daily health settings is underexplored. In this work, we aim to build a comprehensive personal health agent that is able to reason about multimodal personal health data from everyday consumer wellness devices and medical records and provide personalized health recommendations. To understand end-users' needs when interacting with such an assistant, we conducted an in-depth analysis of web search and health forum queries, alongside qualitative insights from users and health experts gathered through a user-centered design process. Based on these findings, we identified three major skill categories to fulfill consumer health needs, each of which is supported by a specialist sub-agent in our system: (1) a data science agent that analyzes personal time-series wearable and medical record data, incorporating population-level statistics to provide contextualized numerical health insights, (2) a health domain expert agent that integrates users' health and contextual data to generate accurate, personalized insights based on health domain knowledge, and (3) a health coach agent that synthesizes data insights, drives multi-turn user interactions and interactive goal setting, guiding users using a specified psychological strategy and tracking users' progress. Furthermore, we propose and develop *Personal Health Agent (PHA)*, a multi-agent framework that enables dynamic, personalized interactions to address individual health needs. To evaluate each sub-agent and the multi-agent system, we conducted automated and human evaluations across 10 benchmark tasks, involving more than 7,000 annotations and 1,100 hours of effort from health experts and end-users. Our work represents the most comprehensive evaluation of a health agent to date and establishes a strong foundation towards the futuristic vision of a personal health agent accessible to everyone.[#]

1. Introduction

Rapid advancements in large language models (LLMs) have driven transformative changes across numerous domains, including health and medicine. With their ability to embed extensive medical knowledge and perform complex reasoning, LLMs have shown promise in assisting with many tasks, such as medical documentation (Lee, 2018), clinical diagnosis (McDuff et al., 2023), decision support (Benary et al., 2023), and drug discovery (Gottweis et al., 2025; Vert, 2023). In parallel, the adoption of LLMs in consumer health domains is also accelerating. Examples include virtual health assistants (Khasentino et al., 2025), symptom checkers (Fraser et al., 2023), digital coaches & advisors (Oura Team, 2025; WHOOP, 2023), and clinical examination dialogues (Tu et al., 2024). These applications primarily focus on leveraging LLMs' linguistic and contextual understanding to provide accessible health and wellness insights.

Corresponding author(s): { *aliheydari, xliucs, dmcduff, orson* } *@google.com*
© 2025 Google. All rights reserved.

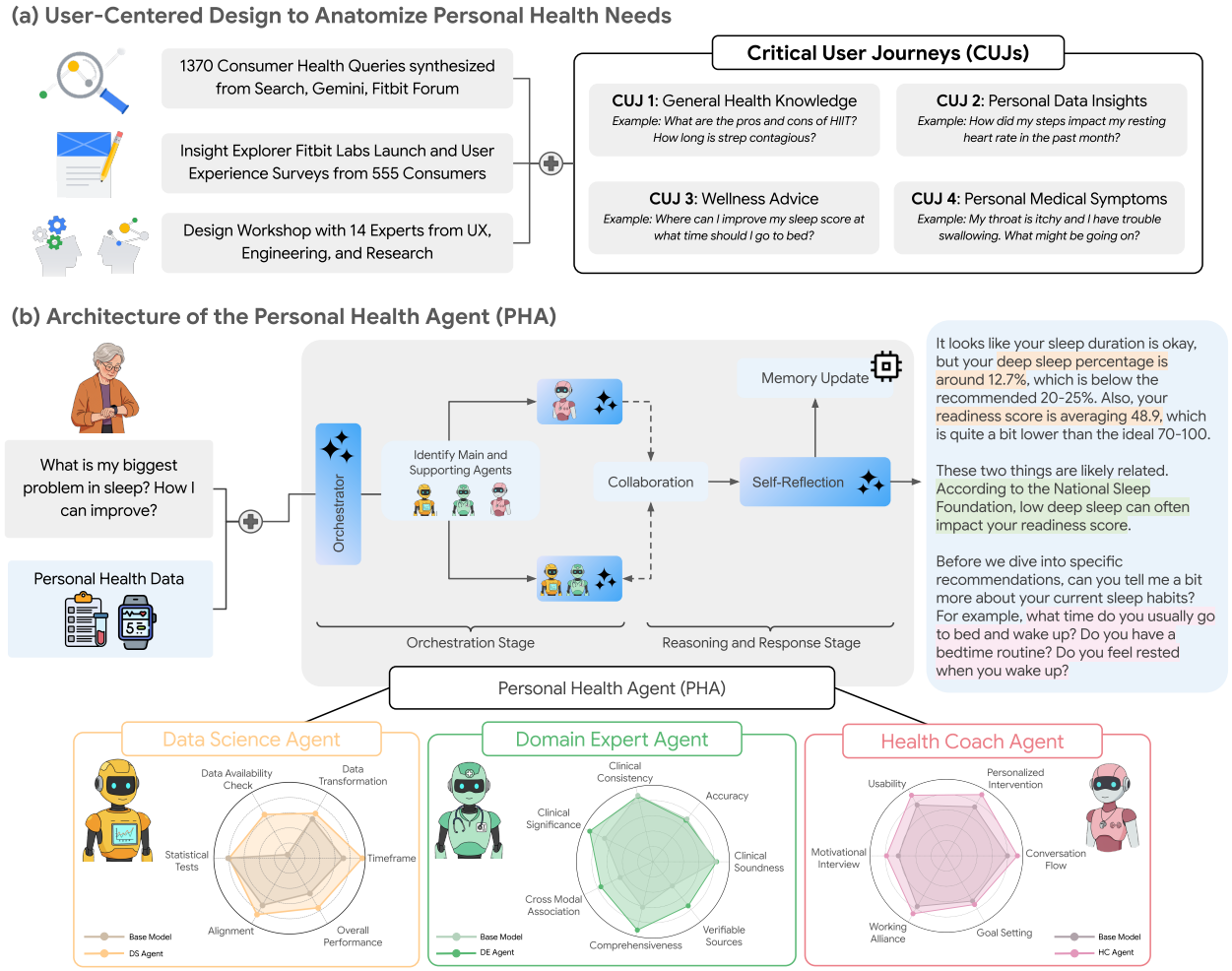
[#] This paper outlines a conceptual framework for research purposes. The PHA and its components are theoretical constructs and should not be considered a description of any specific product, service, or feature currently in development or available to the public.

With the growing adoption of everyday wearable devices, vast amounts of personal health data are generated passively and continuously. These data streams offer a granular, real-time view of personal health behaviors and biomarkers. Signals such as physical activity levels, sleep patterns, heart rate variability (HRV), and a wide range of physiological biomarkers can reflect various aspects of personal health and wellness (Zheng et al., 2024). There is growing evidence that consumer wearable devices can significantly contribute to promoting healthy behaviors (Ringeval et al., 2020). Increasingly, communities across artificial intelligence (AI), human-computer interaction (HCI), health informatics, and other domains have started exploring how LLMs can analyze and interpret such data streams to decode rich information of an individual's health and uncover actionable insights (Khasentino et al., 2025; Merrill et al., 2024a). This creates new opportunities to explore how insights from wearable data, when integrated with personal medical records, could provide personalized health assessments, generate tailored evidence-based wellness recommendations, and encourage behavior change (Jörke et al., 2025; Kim et al., 2019). However, LLM reasoning with complex, numerical time-series data remains a significant challenge (Merrill et al., 2024b) and the prior work falls short of creating comprehensive agentic systems to address a wide range of user queries. For instance, Khasentino et al. (2025) evaluated LLMs on case studies with wearable data, but their analysis lacked a deep assessment of numerical reasoning and focused only on a narrow set of fitness and sleep insights. In a separate study, while Merrill et al. (2024a) tackled a wider range of user queries on numerical data, their work did not evaluate an agent's grasp of domain knowledge or its ability to perform expert-level interactive coaching. These limitations illustrate a necessary shift beyond single-purpose health agents. Given the critical role of integrated individual information for personal health journeys (Graham et al., 2024) and building upon the multifaceted development of LLMs' capabilities, we envision that **the era of intelligent personal health agents is arriving**.

Prior research has found that individual needs for daily health and wellness support are highly diverse and a range of capabilities are required (Merrill et al., 2024a; Srinivas et al., 2025). For instance, a closed-ended query of "On average, how many hours have I been sleeping this last month?" requires different skills versus an open-ended query of "what can I do to improve my sleep quality?". Therefore, personal health agents need to be equipped with a broad set of capabilities to handle various requests from users. In this work, we aim to build a comprehensive personal health agent that integrates wearable sensor data and medical records with LLMs to provide dynamic, personalized interactions capable of addressing a broad range of individual health needs.

We adopted a user-centered design process to architect a health agent that covers a wide range of general wellness, fitness, and medical topics. To better understand users' specific needs, we collected and analyzed real-world data from multiple sources across end-user online queries and discussion, survey results, as well as a expert design workshop. Synthesizing these data revealed four critical categories of user queries (see Supplemental Table S1), including (1) general health and wellness knowledge, (2) personal data, (3) wellness advice, and (4) personal health symptoms.

These four categories highlight several important capabilities that a personal health agent needs to be equipped with: analyzing personal data, interpreting the data in health contexts, and providing personalized actionable suggestions to users. This sheds light on the three synergistic key roles of our personal health agent, each empowered by an advanced LLM sub-agent with specific skills or roles: The *Data Science Agent* (DS Agent) analyzes the user's personal health data from wearables (e.g., Fitbit) alongside population-level time-series data to derive numerical insights, such as estimating changes in running speed from workout logs (e.g., *"Has my running gotten faster since last month?"*). The *Domain Expert Agent* (DE Agent) draws on personal medical records, wearable data, and health knowledge bases to provide domain-specific and contextualized interpretations, such as explaining specific biomarkers or general health conditions (e.g., "Is a blood pressure of 137 over 83 fine?") or comparing a user's data to general population statistics. Finally, the *Health Coach Agent* (HC Agent)



(c) Evaluation Studies for Individual Sub-Agents and the Entire Multi-Agent PHA System

	Assessment Goal	Evaluation	Scale	Human Hours
DS.1	The agent's ability to design a **statistical analysis** plan.	Human: Data Scientists; Auto Eval*	354 Plans x 2 Agent Conds (708 Results)	47
DS.2	The agent's ability to **generate code** to execute the analysis plan.	Human: Software Engineers; Auto Eval*	173 Tests (25 Queries) x 2 Agent Conds (346 Results)	75
DE. 1	The agent's **medical knowledge** and its ability to **reason over case studies** mirroring real-world scenarios.	Auto Eval	2145 MCQs x 2 Agent Conds (4290 Results)	-
DE. 2	The agent's ability to **generate a differential diagnosis** when provided user symptoms.	Auto Eval	1511 User Symptoms x 2 Agent Conds (3022 Diagnosis)	-
DE. 3	The agent's ability to **personalize answers to medical questions** based on context and information about the user	Human: Generalists (End-users)	50 Queries x 2 Agent Conds x 17 Raters (1700 Annotations)	13
DE. 4	The agent's ability to **reason and interpret multi-modal health data** (wearables, lab results, demographics, health surveys)	Human: Experts (MD physicians)	30 Personas x 2 Agent Conds x 5 Raters (300 Annotations)	77
HC.1	The agent's ability to **conduct coaching conversations** from *end-users'* perspectives.	Human: Generalists (End-users)	31 Raters (Multi-turn) x 2 Agent Conds (62 Annotations)	48
HC.2	The agent's ability to **conduct coaching conversations** from *health coach experts'* perspectives.	Human: Experts (Health Coaches)	31 Interactions x 2 Agent Conds x 3 Raters (186 Annotations)	90
P.1	The multi-agent's ability to **synthesize personal data (DS), health domain knowledge (DE), and coaching (HC) capability** in a conversation about achieving personal health goals, from *end-users'* perspectives	Human: Generalists (End-users)	50 Personas x 3 Agent Conds x 20 Raters (3000 Annotations)	500
P.2	The multi-agent's ability to **synthesize personal data (DS), health domain knowledge (DE), and coaching (HC) capability** in a conversation about achieving personal health goals, from *coach experts'* perspectives.	Human: Experts (Health Coaches)	50 Personas x 3 Agent Conds x 5 Raters (750 Annotations)	270

Figure 1 | Overview of The Personal Health Agent. **(a)** We present the anatomy of personal health needs with a user-centered design process. **(b)** We then design an multi-agent system, powered by state-of-the-art large language models, consisting of specialist Data Science (DS), health Domain Expert (DE), and Health Coach (HC) agents to address diverse consumer health needs. "Base model" refers to the Gemini 2.0 family. **(c)** Our work represents the most comprehensive evaluation of a health agent to date (>7,000 annotations and 1,100 hours of effort on 10 benchmarks).

applies evidence-based psychological strategies, like motivational interviewing (Basar et al., 2024), to help users set appropriate goals, identify barriers, and develop personalized plans to foster lasting behavior change.

The roles of these sub-agents are not independent. A wide range of end-user queries naturally involve more than one agent. For instance, the query "I want to understand my sleep data last week and know how I can improve my sleep" would require the DS Agent to analyze sleep and other related wearable data and present the results, the DE Agent to provide evidence-based sleep knowledge and the HC Agent to engage in dialog with the user and provide personalized sleep advice. Therefore, to provide a comprehensive user experience, we developed the ***Personal Health Agent (PHA)***, a multi-agent system composed of an orchestrator and the three specialized sub-agents, as shown in Figure 1(a). By design, our PHA integrates capabilities that span data science, health domain expertise, and coaching, *intentionally blurring the lines between consumer wellness and clinical knowledge domains*. This approach is not intended to supplant human health providers, but to explore the technological frontier of consumer empowerment, envisioning how a unified agent can provide more holistic, actionable health insights.*

Our PHA enables coordinated, context-aware interactions across agents. Drawing inspiration from a range of multi-agent collaboration strategies, such as voting (Chen et al., 2024; Wang et al., 2022), iteration (Tang et al., 2024), debate (Du et al., 2024), and hierarchical team formation (Kim et al., 2024), we propose a set of design principles for multi-agent collaboration for health applications, leveraging the advantages of different individual agent capabilities, and design our PHA architecture with iterative orchestration and collaboration. This multi-agent framework enables dynamic collaboration across the three sub-agents and enables personalized interactions with end-users, seamlessly combining data analysis, domain expert knowledge, and health coaching to support a broad range of individual health needs.

To validate our system, we developed a holistic evaluation framework (Figure 1(b)) that assesses performance at two distinct levels: individual sub-agent capabilities and the integrated multi-agent system's overall efficacy. At the individual level, we evaluated each agent's core competencies: the DS Agent was benchmarked on its ability to generate robust analysis plans (Section 4); the DE Agent on its capacity for evidence-based reasoning and providing factual knowledge (Section 5); and the HC Agent through user-centered studies involving both end-users and experts to assess its coaching effectiveness (Section 6). At the system level, we evaluated the integrated multi-agent PHA framework through comprehensive studies with both end-users and health experts, using open-ended, multi-modal conversations that reflect diverse health scenarios (Section 7). In total, our evaluation comprises 10 benchmark tasks and draws upon over 7,000 human annotations across 1,120 hours of human effort (expert: 559 hours, end-users: 561 hours), validating the advantage of each specialized sub-agent and demonstrating the promising real-world applicability of our integrated multi-agent health agent.

By providing a comprehensive multi-agent architecture and a holistic evaluation framework, our work establishes a concrete foundation for the future development of accessible personal health agents to help improve individuals' everyday health and well-being.

2. A User-Centered Design of Personal Health Agent Requirements

We adopted a user-centered approach to identify and categorize end-users' needs for a personal health agent. Our process integrated insights from three layers of complementary data collection: *real-world user queries, targeted survey responses, and expert synthesis*. Together, these layers form a

*It is noteworthy that our user-centered design process and the PHA system development was for academic and foundational research purposes only, and is not tied to the development of any current or planned product.

coherent pipeline: we began by capturing broad, naturalistic user intent through consumer health queries; we then deepened this understanding through structured survey data that elicited users' goals and satisfaction. Lastly, we conducted expert workshops to distill and organize these findings into actionable requirements.

2.1. User-Centric Data

The collection and interpretation of user-centric data included three aspects:

Consumer Health Queries. We compiled and analyzed consumer health queries from Google Search, Gemini, the Fitbit community forum, from January 2024 to February 2025. In addition, in October 2024 through February 2025, we launched Insights Explorer through Fitbit Labs (Fitbit Community, 2024). Participation was entirely voluntary and required users to explicitly opt-in and provide in-app consent for their data to be used for research and development. As part of this consented experience, participants were able to ask health queries and receive responses from a prototype generative AI conversational agent. We composed a set of 1370 de-identified user queries that represent a wide range of questions regarding daily health and wellness from real-world users. Supplemental Table S1 lists out some examples of these queries.

Fitbit Labs Survey Data. To enrich the query data, we also collected survey data from participants (N=555) in the Insights Explorer Fitbit Labs to ask about their personal health goals and sentiments when interacting with the agent. Specifically, the survey includes questions about user needs (e.g., "What are you hoping to achieve by using [the prototype]?", "Which of the following aspects of health and wellness, if any, did you explore in [the prototype]?"), user experience and satisfaction (e.g., "Indicate any aspects of [the prototype] that have / have not met your expectations") , engagement and retention (e.g., " What was the main functionality or benefits that you continued using [the prototype] for?"), and other user characteristics (e.g., prior experience with wearable sensing, AI, and conversational agents). Supplemental Table S2 lists out details of these lab surveys.

Expert Workshop. To complement users' perspective, we further conducted a one-hour critical user journey workshop with experts across user experience, product, research and engineering (N=14) to categorize the data collected. The workshop procedure focused on summarizing the user journey goals and the specific associated tasks. After collecting multiple experts' perspectives, we adopted a standard thematic analysis process (Maguire and Delahunt, 2017). Two researchers independently coded the same subset of these queries and all survey data to jointly develop a codebook of hierarchical user journeys (categories and sub-categories). The same researchers then categorized the rest of the user queries into these categories and expanded the codebook as needed.

2.2. Major Categories of Use-Cases with Personal Health Agent

Consolidating results from user queries, surveys, and the expert workshop, we identified four representative categories of critical user journeys (CUJ) for a personal health agent.

CUJ 1 **General health knowledge**: These queries focus on factual knowledge related to broad health topics across medical, fitness, and general wellness areas, where users aim to expand their understanding of specific health concepts and improve their overall health knowledge. This category includes questions aimed at discovering and comparing health-related facts, exploring the pros and cons of different treatments or behaviors, and staying up-to-date with recent health news.

Examples: (i) *"How long is strep contagious?"* (ii) *"What are the pros and cons of HIIT (high intensity interval training)?"* (iii) *"What is the difference between HRV (heart rate variability) & RHR (resting heart rate)?"*

CUJ 2 **Personal data insights**: This category includes queries where users seek to understand their own health data, whether collected through wearable devices or available in their personal medical records (e.g., lab results). These questions often involve interpreting patterns in their health metrics, exploring the relationship between data aspects, comparing their measurements against clinical guidelines, and identifying meaningful changes.

Examples: (i) *"What is my average resting heart rate on days when my sleep score is above 86?"* (ii) *"How did my steps impact my resting heart rate in the past month?"* (iii) *"Can you review my HRV data from this month and highlight any unusual occurrences?"*

CUJ 3 **Wellness advice**: In this category, users mainly focus on seeking actionable insights and health advice, either general or data-specific, to enhance their wellness. Queries include requests for general wellness tips, tailored advice based on personal data, and support in identifying potential areas of improvement. Some queries also involve asking guidance on setting or adjusting goals, creating plans to achieve those goals, overcoming obstacles, and tracking their progress.

Examples: (i) *"To improve my sleep score, what time should I go to bed?"* (ii) *"It's raining. Adjust my workout to be indoors today."* (iii) *"What are some of the things I have done that have the biggest impact on my blood pressure?"*

CUJ 4 **Personal medical symptoms**: This category involves asking about specific symptoms or assessing potential causes that are related to users' personal experience. Some queries require back-and-forth conversations to collect additional information from users.*

Examples: (i) *"My throat is itchy and I have trouble swallowing. What might be going on?"* (ii) *"How can I tell if I have high blood pressure?"* (iii) *"I just threw up blood. What might be going on?"*

It is worth noting that these categories are not mutually exclusive at the level of an individual query or conversation. For example, a query about personal data can be embedded in another query for wellness advice. Furthermore, these do not exhaustively cover all user needs but rather encompasses a wide range of needs and we exclude other queries that are less related to health (e.g., privacy, safety, social relationships). Supplemental Table S1 provides a summary of these categories.

3. Personal Health Agent and Evaluation Framework

Our analysis of user needs identified four CUJs, spanning data interpretation, knowledge seeking, wellness advice, and symptom assessment. Satisfying these diverse requirements demands more than a single monolithic system and calls for a set of specialized, complementary capabilities. Our approach is therefore grounded in both the cognitive theory of modular cognition (Fodor, 1983), which posits that the mind uses specialized subsystems for different tasks, and the empirical success of human health teams, where experts with complementary skills collaborate. Inspired by this, we deconstruct the problem space and propose a multi-agent personal health assistant that brings together three core roles, each embodied by a specialized sub-agent: data analysis, health domain expertise, and behavior coaching, working in concert to emulate a professional health support team.

3.1. User-Centered Agent Design

*Due to the high-stakes nature and ethical considerations, in the real-world, we envision that personal health agents would focus on providing end-users with contextual information and help connect them with real-world clinical resources. The goal of these agents is to empower rather than replace clinicians. In this work, we address the research aspects of this CUJ.



Data Science (DS) Agent: This agent mainly focuses on addressing user queries about personal data. Existing research has explored LLMs' ability to answer data science questions (Cheng et al., 2023; Gu et al., 2024b; Hong et al., 2024; Hu et al., 2024a; Li et al., 2024; Merrill et al., 2024a; Wu et al., 2024a; Yin et al., 2023; Zhang et al., 2025, 2024), but their capacity to interpret open-ended, personalized data inquiries and perform statistically sound data analyses remains unclear, and there is a lack of open benchmarks assessing their reliability in this area. We develop the first personal open-ended data science agent capable of analyzing both personal and population-level time-series wearable data to provide numerical health insights. Our agent enables users to ask diverse, open-ended, and personalized questions about their health data and provide meaningful, data-driven responses, filling a critical gap in personalized data analysis. We also open-source an evaluation benchmark for these capabilities. Example conversations between a user and the DS agent are shown in **Supplemental Figure S3**.



Domain Expert (DE) Agent: This agent aims to answer any queries that are related to medical knowledge or daily symptoms. While prior work has evaluated LLMs' medical knowledge (Nori et al., 2023, 2024; Singhal et al., 2023), many essential skills of an ideal domain health expert are underexplored, such as interpreting data within contexts, e.g., population and environmental factors, and performing summarization and reasoning across data modalities from wearables and medical records. We build a domain expert agent equipped with a suite of tools to address these challenges. The agent integrates users' wearable data, medical records, and contextual data to generate accurate, relevant responses to complex health inquiries. It advances the state-of-the-art LLMs by providing nuanced multi-modal interpretations that consider the broader context of personal health. Example conversations with the DE agent are shown in **Supplemental Figure S4**.



Health Coach (HC) Agent: The HC agent aims to provide personalized health and wellness advice, and to help users set up goals and motivations via conversations. Although previous studies have initiated explorations into the application of AI in personal coaching (Jörke et al., 2024; Mantena et al., 2025; Mercado et al., 2023), our agent is the first to support open-ended coaching conversations across a wide range of general wellness topics. Building on coaching expert insights, we significantly enhance the agent's coaching skills, including motivational interviewing and personalized health recommendations. This agent adopts a modular structure to optimize the conversation flow, balancing information gathering, active listening, and recommendation delivery to support a smooth and effective coaching experience. Example multi-turn conversations with the HC agent can be found in **Supplemental Figure S9-Figure S10**.



Personal Health Agent (PHA) – Multi-Agent Collaboration: The three sub-agents have complementary skill sets and often need to collaborate together to provide appropriate support regarding a user's query. A central orchestrator manages this collaboration. It receives the user's initial query, parses it, and determines which agent(s) are best suited to respond. It decomposes the query into sub-tasks, routes them to the appropriate sub-agents, synthesizes the responses into a cohesive, user-friendly answer, and then reflect and iterate on it prior to responding to users (Tran et al., 2025). **Supplemental Figure S11-Figure S14** demonstrates a multi-turn conversation example with PHA.

3.2. Agent Evaluation Framework

In order to systematically evaluate our agent framework and identify the strengths and gaps in the design of each component, we created the evaluation framework shown in Figure 1(b). We took a mixed-methods approach and conducted a set of studies that examine a comprehensive set of factors:

1. **Interaction Turn:** Single and multi-turn interaction.

2. **Data Modalities:** Agent processing with both single- and multi-modality data.
3. **Evaluation Setup:** Automatic and human evaluations of conversations.
4. **Human Perspective:** Health experts' and end-users' perspectives on multi-turn conversations.

This multi-faceted approach provides a holistic understanding of the AI agent's performance and is the most comprehensive evaluation of a consumer AI health agent to date.

Specifically, our framework delineates and assesses the core capabilities of each component. The DS Agent is evaluated on its two-stage process of **(DS.1)** generating robust analysis plans and **(DS.2)** translating them into accurate, executable code. The DE Agent is assessed across four key competencies: **(DE.1)** answering common health questions, **(DE.2)** handling diagnostic conversations, **(DE.3)** personalizing answers to medical questions based on context and information about the user, **(DE.4)** interpreting and reasoning with multi-modal health data (wearables, lab results, demographics, health surveys). The HC Agent's capability of providing personalized coaching experience is evaluated from **(HC.1)** its effectiveness from the end-user's perspective and **(HC.2)** its fidelity to human expert coaching principles. After connecting all three sub-agents, the final multi-agent PHA system has the comprehensive set of skills to support various end-user health queries at a human health expert level. It is assessed on its holistic conversational performance from complementary viewpoints of both **(PHA.1)** end-users and **(PHA.2)** experts.

3.3. Choice of Base Language Model

We used the Gemini 2 family (Flash/Pro) of models (The Gemini Team, 2023) for our experiments. Gemini has demonstrated strong performance on medical tasks (Khasentino et al., 2025; Saab et al., 2024; Yang et al., 2024). In this work, our focus is not to train a new LLM but rather to design and construct the agentic system around a given base LLM. By selecting Gemini as the base model we could ensure that (i) there was no training data contamination, and (ii) no health data from real participants would be logged by the model and used for training future models. While existing contamination was not a significant concern, as a majority of our analyses involve novel datasets that are unlikely to be part of any LLM training dataset, choosing Gemini allowed us to verify that any testing of the models did not lead to data entering a future training set. We note that our design of sub-agents and the agent teams are agnostic to the base LLM. We expect that our proposed solutions will be generalizable to other LLMs.

3.4. Real-World Dataset for Evaluation

To ground our agent evaluation with real-world validity, we evaluated our framework's applicability and robustness against the *Wearables for Metabolic Health (WEAR-ME) study*, a large-scale (N=1165) effort to investigate metabolic health (Metwally et al., 2025). The overall study design is outlined in Figure 2(a). As part of WEAR-ME, consented participants were asked to link their Fitbit account to Google Health Studies (GHS), authorizing the collection of their wearable data for the study's duration and for up to three months before they joined. Once enrolled, participants were asked to (i) complete questionnaires covering demographics, health history, and personal health insights, (ii) schedule a blood draw with a Quest Patient Service Center, and (iii) wear their Fitbit or Pixel Watch throughout the day and night (for at least 3 out of every 4 days). In total, WEAR-ME contains multimodal data from each participant on their questionnaire answers, Fitbit records, and blood test results. More details of the dataset are described in Supplemental Section B.2. The WEAR-ME study was conducted in full compliance with privacy regulations, with the study approved by the Advarra Institutional Review Board (IRB, #Pro 00074093). All participants provided informed e-consent for their data to be used in research and publications.

Throughout our evaluation, we leveraged the WEAR-ME dataset for various evaluation tasks (see tasks **DS.1**, **DE.3**, **DE.4**, **HC.1**, **HC.2**, **PHA.1**, **PHA.2** in Figure 1). As introduced in evaluation details in later



Figure 2 | Overview of the wearables and metabolic health study (WEAR-ME) study. (a) The WEAR-ME study consists of Fitbit users who opted-in to participate in our study. Those who consented and enrolled in the WEAR-ME study were then asked to visit one of the Quest centers for blood biochemistry data collection. Each participant contains multimodal data from Fitbit, blood test results, and questionnaires. **(b)** For our human evaluations, we selected a subset of the WEAR-ME data based on prevalent health conditions present among the WEAR-ME participants. We identified 10 common profiles, and selected five individuals at random from those profiles. More details are introduced in Supplemental Section B.

sections, we contextualize our end-user and expert evaluations on a subset cohort of WEAR-ME for real-world validity. Working with clinicians, we identified the 10 most common health profiles among the participants and then randomly selected five individuals representing each of those conditions (see the specific 10 profiles and persona examples in Supplemental Section B.2). These 50 user personas were used to generate evaluation samples for our agents' multiple evaluation tasks (e.g., **Section 5.3.3, Section 5.3.4** for the DE agent, and **Section 7.3.1, Section 7.3.2** for the PHA). The construction of the evaluation cohort is depicted in Figure 2(b).

4. Data Science Agent

Data science is the extraction of knowledge from information and enables the identification of patterns and prediction of outcomes (Dhar, 2013). The increasing digitization of healthcare has resulted in an unprecedented volume of personal health data, including electronic health records, genomics, medical imaging, and wearable sensor data. While prior work has studied and evaluated data science agents capable of executing code given clear, precise instructions (e.g., "Calculate the mean and standard deviation of temperature value in the 'Mar.2019' column") (Gu et al., 2025a; Guo et al., 2024; Hu et al., 2024b), it remains unclear whether LLM-based agents can interpret, decompose, and operationalize *open-ended*, *underspecified*, and *personalized* data inquiries (e.g., "Am I sleeping well?") into statistically sound analyses.

This capability is especially critical in high-stakes domains like personal health, where flawed analytical decisions can lead to misleading or even harmful conclusions (Simonsohn et al., 2020; Steegen et al., 2016b). However, despite its importance, there is currently no systematic benchmark for evaluating analysis planning in the context of personal health. To address this gap, we focus on the **Data Science (DS) Agent**'s ability to perform robust statistical analyses on personal health data and deliver valid and actionable numerical insights.

4.1. Data Science Agent Capabilities

The specific statistical methods that are necessary to answer a given user query can vary. For example, "Does performing activities give me better sleep?" and "How much exercise do I get compared to my age group?" require different reasoning steps. The DS Agent needs to handle diverse user queries, many of which are often underspecified and ambiguous, i.e., multiple valid approaches or perspectives in answering the question (Liu et al., 2020a, 2019; Steegen et al., 2016a). Nevertheless, an ideal DS Agent should be able to perform a logical and rigorous analysis that aligns with at least one of the core intents of the query (i.e., the user will get a satisfactory answer from the analysis) and provide numerically accurate and statistically robust results from which the user can derive confident conclusions (e.g., not claiming significant trends from an insufficient number of data points).

We focus on a set of capabilities for the data science agent that are useful, yet tractable to evaluate (Gu et al., 2024b; Merrill et al., 2024a; Wu et al., 2024b). These capabilities are informed by the types of user interactions and data that are expected in the context of the CUJs derived in Section 2.2. Broadly, we define the capability of the DS Agent to address queries that involve analysis of numerical data (e.g., wearable streams, medical records), with the goal as obtaining statistical outcomes that align with the query intent and context surrounding those statistics, such as timeframe, type of statistics, assumptions, etc. Specifically, the capabilities of the agent are to:

DS1. **Decompose An Open-Ended Query into a Structured Analysis Plan:** For broad or ambiguous queries often posed by end-users without data science expertise, the DS Agent should translate them into appropriately specific, unambiguous, tractable queries and analysis plans by:

a) Identifying conceptual variables (Jun et al., 2022): Recognizes the concepts in the

question that can and should be represented in the data (e.g., representing fitness as average daily steps and resting heart rate for a query like *"Am I getting more fit over the past few months?"*).

b) Formulating the correct and exact data transformation steps to operationalize the variables across multiple tables of user data and population-level data tables. This involves the following optional steps:
   i) Filtering relevant subsets of data (e.g., filter for activities in the last 3 months),
   ii) Cleaning messy data (e.g., handling missing or misrecorded wearable data)
   iii) Joining multiple tables (e.g., joining the activities data table with the daily summary data table on the same day),
   iv) Aggregating data (e.g., getting average steps and resting heart rate grouped by month).

c) Recognizing, highlighting or addressing missing/insufficient data to prevent unreliable conclusions.

d) Applying appropriate statistical tests and calculations based on the user's intent.

Note: Given the open-ended nature of user queries, multiple interpretations and answers may be valid. Accordingly, in one single interaction, our DS Agent is designed to provide at least one appropriate response, rather than an exhaustive enumeration of all possibilities. Users can have follow-up conversations to explore other options with the agent. In the case where the data in user queries is not available, the agent might check the availability of the data with users and request data from users, if needed.

DS2. **Compute accurate numerical results through code generation, execution, and feedback:** Based on a precisely specified analysis plan (i.e., one without any ambiguity for alternative methods and models Liu et al. (2020b)), the agent should be able to generate, execute, and debug code that follows the exact specifications of the analysis plan, such that it can execute bug-free code and compute accurate results.

4.2. Data Science Agent Architecture

Based on the delineated capabilities, the DS Agent's architecture is designed as a modular, two-stage pipeline: (1) Analysis Plan Generation and (2) Code Generation with Iterative Execution (Gu et al., 2024a). This decomposition is critical for mitigating the risk of factual inaccuracies and logical errors common in end-to-end generation for complex data tasks (Jiang et al., 2024). It separates the semantic reasoning of "what to do" from the syntactic task of "how to do it."

The first stage translates an open-ended natural language query into a precise, structured analysis plan. To ground its reasoning, the planner is provided with a rich contextual summary of the available data, including table schemas, column descriptions with data types and semantic tags, and summary statistics (Dibia, 2023). The output is a structured natural language paragraph that explicitly outlines the conceptual variables, required data transformations (e.g., filters, joins, aggregations), and the chosen statistical model, directly instantiating the components listed in **Section 4.1**.

The second stage receives this structured plan and is tasked with generating executable code. This focus on a well-defined specification allows the LLM to leverage its strong code generation capabilities (Jiang et al., 2024). The generated code is then executed in a sandboxed environment. We employ an iterative refinement loop: if execution fails due to syntactic errors or runtime exceptions, the agent re-invokes the stage with the original query, plan, and the error message as feedback. This self-correction cycle, which has been shown to improve robustness (Quoc et al., 2024), continues until bug-free execution is achieved and a numerical result is produced.

Figure 3(a) illustrates the architecture enabling these functions in the DS Agent. The details of agent

prompt and architecture are included in **Supplemental Section C.3**.

4.3. Data Science Agent Evaluation

To ensure our DS Agent produces statistically robust and numerically accurate answers, we independently evaluate the two core components in two tasks.

4.3.1. Task 1: Analysis Plan Generation

Evaluation Goal. Given an open-ended query involving data analysis (e.g., *"Am I getting more fit recently?"*), we measure the agent's ability to decompose and/or expand the query into a precise and robust statistical analysis plan. The plan should align with the query's intent and appropriately involve the analysis of personal health data.

In this paper, we focus on the data analysis on the wearable data from the WEAR-ME dataset (Section 3.4). In particular, the DS Agent is given access to the following data tables for each individual user: (1) a daily summary table with daily activity and sleep wearable data, (2) an activities table where each row represents a recorded physical activity (e.g., running, biking, swimming etc.), and (3) a population summary table that provides daily wearable percentiles stratified by gender and age group. We include the full data schema of all tables used by the DS Agent in **Supplemental Section C.2**.

Evaluation Setup. An analysis plan involves many decisions that contribute to the robustness of the final result, such as handling missing data, choosing a statistical model, and operationalizing constructs (Liu et al., 2020b,c; Steegen et al., 2016b). Evaluating these decisions with human experts is prohibitively expensive. To scalably evaluate a wide range of analysis approaches, we develop another LLM as an autorater (Chiang and Lee, 2023; Zheng et al., 2023) tested against human expert raters to ensure its reliability.

Evaluation Rubrics and Point-Detection Metrics. To enable this, we first developed a detailed analysis plan rubric in consultation with a panel of four data science experts. The rubric is designed for objective assessment across six key dimensions:

1. **Timeframe:** Does the analysis consider a specific timeframe (e.g., past month, past 6 months) when it helps personalize and align with user intent?
2. **Data Transforms:** Does the approach correctly reference existing columns and apply any necessary data transformations, without logical errors, in a way that others can reliably reproduce?
3. **Data Sufficiency:** Does the approach reasonably and clearly assess whether there is enough data, both in terms of available rows and non-missing cells within columns, to support the intended transformations, calculations, and statistical analyses?
4. **Summary Statistics:** Does the approach appropriately calculate summary statistics where helpful, and are the chosen statistics reasonable given the context and question being answered?
5. **Statistical Tests:** Does the approach appropriately apply statistical tests when helpful, with clear, reproducible methods, reasonable test choices, and proper consideration of distributional assumptions?
6. **Overall Alignment:** Finally, given the evaluation of the above dimensions, does the approach align with the user's intent?

Our final rubric consists of 26 items arranged in a conditional hierarchy. This structure means some items only apply if a preceding one is satisfied; for example, we only assess the appropriateness of a statistical test if the initial hypotheses are sound. To quantify quality, we used a point-deduction system where errors or omissions lower the final score. Because of this hierarchy, a maximum of 16 points can be deducted in any single evaluation. The detailed rubric development process and point system are provided in **Supplemental Section C.4.1**.

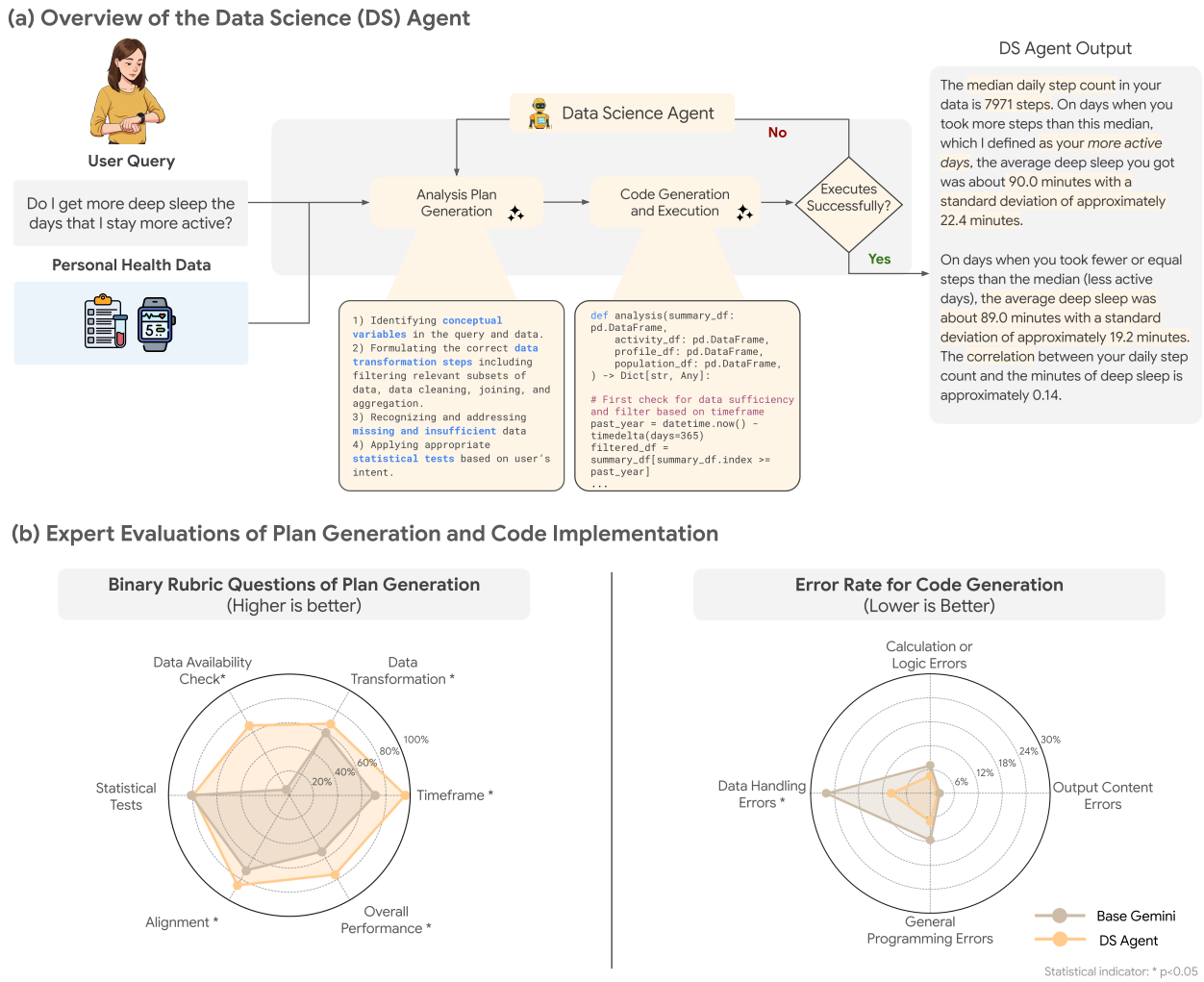


Figure 3 | Schematic and evaluation results of the *Data Science (DS)* Agent. (a) The Architecture of the DS Agent, which first generates a detailed, natural language statistical analysis plan from a user's query and a data schema summary (see Supplemental Section C.3 for prompting details). Subsequently, it produces and executes code to implement this plan. An iterative self-correction loop allows the agent to revise the code based on execution errors, the original query, and the analysis plan. **(b)** Results of the Data Science Agent's performance evaluation, demonstrating significant improvements over the base Gemini model in both analysis plan quality and code generation pass rates. Supplemental Section C.4 summarizes evaluation details and rubrics.

Benchmark Dataset of Query-Analysis Plan Pairs. Evaluating the quality of generated analysis plans and developing a scalable autorater requires a reliable benchmark. To this end, we created an expert-rated dataset of 141 unique query–analysis plan pairs. These pairs were generated using various LLMs (Gemini 1.5-2.0, GPT-4o) across 100 distinct health queries (sampled from the end-user health query set collected in Section 2.1). 10 expert data scientists (8 males, 2 females, aged 25 - 45, all Ph.D.s in computer science, with 5 to 20 years of experience) spent 47 hours in total and provided 354 approach annotations based on the rubrics defined above (each query–analysis plan pair was annotated by 2-4 experts), leading to a total of 6157 rubric item annotations. This expert-rated dataset demonstrated a moderate level of inter-rater reliability (IRR), with an average agreement of 80.2% and a Bennett's S of 0.622 (Bennett et al., 1954). Detailed statistics can be found in **Supplemental Table S5**.

Development of a Reliable LLM Autorater. Using this human-rated dataset, we developed and validated an autorater based on Gemini 2.0 Flash. The dataset was split into development and test sets, with the test set comprising 96 evaluations of 25 unique queries. The development set with 75 queries was then used to train the autorater (the specific process is detailed in **Suppmental Section C.5**). Our final autorater achieved an Intraclass Correlation Coefficient (ICC(3,1)) of 0.838 when compared to the scores assigned by human experts, indicating high reliability for scaled evaluation (Zheng et al., 2023).

We then assess the quality of the DS Agent generated analysis plans using this developed autorater, and compare them to the baseline Gemini 2.0 Flash model. The plan generation prompt templates for both DS Agent and the baseline are listed in **Supplemental Section C.3**.

Evaluation Results. The results presented in the left of **Figure 3(b)**. The DS Agent demonstrated a statistically significant improvement in analysis plan quality compared to the baseline Gemini model. In general, the DS Agent achieved an average score of 75.6±1.4% (mean±s.e., meaning 24.4% of the points were deducted according to the rubric), marking a substantial increase over the baseline's 53.7±1.8% ($p < 0.001, r = 0.925$, Wilcoxon signed-rank tests with rank-biserial correlation coefficient r as the effect size). Most notably, the agent demonstrated a transformative improvement on the Data Availability Check, rising from a near-failure rate of 5.3 ± 1.6% to 66.5 ± 3.7%. The base model's deficiencies in detecting missingness and artifacts are well-documented limitations of current large language models (Gu et al., 2025b). Its ability to correctly identify the analytical Timeframe was also significantly enhanced, achieving a performance at 96.0±1.9% ($p < 0.001, r = 0.907$). Significant, albeit more moderate, gains were also observed for Data Transforms ($p = 0.002, r = 0.512$) and Alignment with the user's query ($p = 0.011, r = 0.467$).

Conversely, no significant difference was found for Statistical Tests (p=0.630), a category where the baseline model already performed strongly (81.3%). These findings suggest that while the agent's architecture provides a distinct advantage in planning and grounding an analysis in practical constraints, the selection of standard statistical methods is a task for which the foundational model is already well-optimized. Full statistical details are available in **Supplemental Table S6**.

Overall, these results validate our architectural design for the analysis plan generation part. By first generating an analysis plan explicitly with a structured process (**Section 4.1**), the DS Agent outperforms the baseline in bridging the gap between a user's open-ended, underspecified queries and the concrete plan. This initial planning step is an essential prerequisite for the next step of analysis plan implementation.

4.3.2. Task 2: Code Generation and Execution

Evaluation Goal. Given a precisely specified analysis plan from Task 1, this task evaluates the agent's ability to translate that plan into code that is not merely executable, but also functionally correct and robust. The primary goal is to measure the implementation's fidelity to the plan and the numerical

accuracy of its results, especially its ability to handle key data science challenges. This includes managing edge cases (e.g., limited data, complex data transformations) and provide a correct answer to the user's question.

Evaluation Setup. To objectively evaluate the code, we manually converted the analysis plan into a precise function header and doc string to the agent following prior work (Chen et al., 2021). The code was written in Python in a sandbox environment with libraries such as numpy, pandas, and scipy. The details of implementation are included in the **Supplemental Section C.3**.

In particular, a group of 7 data scientists (5 males, 2 females, aged 22 - 45, with 5 to 10 years of experience) were recruited for the code generation evaluation task. They were provided with 25 query-approach pairs with high ratings sampled from **Section 4.3.1** and were asked to formulate each of them as a coding problem with a function header and doc string (clear return output structure). Next, experts were asked to write the function implementation and associated test cases as unit tests. Going beyond simply checking for executable code, these unit tests were designed to rigorously assess the correctness and robustness of the underlying logic. They evaluated the system's ability to handle key analysis cases, including managing limited data, performing complex data transformations, applying appropriate statistical methods, and ultimately, providing an accurate answer to the user's question. In total, we collected code detailing comprehensive test cases for 173 unit tests from these 25 query-approach pairs, with the number of tests per pair varying based on the complexity of the query (75 hours of human effort). Examples of unit tests are shown in **Supplemental Section C.4.2**. Our DS Agent and the base Gemini 2.0 Flash model were assessed for their code generation capabilities using these 25 test suites.

The pass rate was determined by calculating the ratio of successful test cases (across all 173 tests in 25 suites) to the total number of test cases. If code execution fails, the model will receive error messages and retry up to 5 times.

Evaluation Results. The code generation assessment in the right of **Figure 3(b)** reveals a clear advantage for the DS Agent, which achieved a 75.5±3.3% pass rate (i.e., error rate of 24.3%) on its first attempt, significantly outperforming the 58.4±3.7% pass rate (error rate of 41.6%) of the Base Gemini model (χ^2 = 26.3, $p < 0.001$, McNemar's Test). This demonstrates a marked improvement in generating functionally correct code in a single shot. The agent's performance further improves to 79.0% after five trials, confirming an effective, albeit modest, capacity for iterative self-correction. This indicates that while the primary performance gain stems from the agent's superior initial code generation, the iterative process further offers a valuable mechanism for incremental refinement. This dual capability positions the agent as a reliable and practical tool for automated code generation tasks.

More specifically, the DS Agent substantially outperforms the Base Gemini model in reducing critical code generation errors. The most substantial improvement was observed in data handling errors, which dropped from 25.4±3.3% to 11.0±2.4% (χ^2=12.3, $p < 0.001$). Since these errors typically result in unusable code, their reduction indicates improved robustness in data pipeline construction. Moreover, generations free of data handling issues were subsequently evaluated for general programming, calculation/logic, and output content errors. Across all three categories, the DS Agent consistently showed lower error rates, though the differences did not reach statistical significance (e.g., 3.9±1.5% vs. 7.0±2.2% for logic errors, Z=1.15, p = 0.249). Full details can be found in **Supplemental Table S7**.

The DS Agent's superior performance validates the code generation and execution component of our agent architecture. By successfully translating a robust plan into a reliable numerical output, this capability completes the agent's end-to-end workflow with the two critical components, ensuring that

users receive not only a well-reasoned analysis but also a statistically sound and trustworthy answer.

5. Domain Expert Agent

The increasing volume of personal health data (e.g., medical records, wearables data) and rapidly evolving research literature can overwhelm users. This is particularly true for those without clinical expertise, who may struggle to translate such granular data into actionable health insights or care plans. While LLMs demonstrate impressive capabilities of retrieving general knowledge, including in medicine (e.g., Saab et al. (2024)) and personal health (e.g., Khasentino et al. (2025)), their application in health domains, together with data-driven insights from the DE Agent, demands rigorous attention to information reliability and accuracy. A general-purpose LLM, lacking specific medical grounding on authoritative sources, may produce plausible-sounding but ultimately inaccurate or misleading information (Asgari et al., 2025), posing potential risks to users' wellness.

While studies have investigated the medical knowledge of LLMs (e.g., Kim et al. (2025c); Singhal et al. (2023); Xu et al. (2024)), essential capabilities for an ideal LLM-based health domain expert remain underexplored, especially on the aspect of personalizing responses based on contextual information (such as demographics or pre-existing conditions), and reasoning over multimodal data from wearables and medical records to generate summaries and proactive health assessments.

To address these limitations, we designed the **Domain Expert (DE) Agent**, as a research construct, to function as a specialized cognitive layer for health applications. Its purpose is not merely to retrieve information, but to synthesize, contextualize, and validate it against authoritative sources and patient-specific data. By acting as a focused and reliable source of medical expertise, the DE Agent aims to minimize misinformation risk and translate complex data into clear, actionable insights, serving as a critical foundation for our multi-agent design.[†]

5.1. Domain Expert Agent Capabilities

To fulfill this role, the DE Agent needs to move beyond the simple query-response behavior of standard LLMs. It should be equipped with a set of distinct yet interconnected skills that mirror the reasoning of a human clinician. This involves not only accessing factual knowledge but also applying it within the context of an individual's specific health status and history. To formalize this multifaceted expertise, we define and evaluate the agent's core competencies across four foundational capabilities:

DE1. **Accuracy of Health Information:** The DE Agent should provide accurate and reliable medical knowledge spanning diverse personal health and clinical domains, ensuring information is consistent with expert-level understanding.

DE2. **Diagnostic Reasoning Acumen:** The agent can perform differential diagnostic reasoning by generating and ranking plausible diagnoses based on a user's symptoms and relevant medical history conveyed through natural conversations.

DE3. **Contextualization and Personalization Efficacy:** The agent should contextualize and personalize health information and advice, tailoring outputs to a user's specific demographic profile, pre-existing conditions, and health goals to ensure relevance, clarity, and safety.

DE4. **Multimodal Data Synthesis and Reasoning:** The DE Agent should synthesize and reason over heterogeneous data sources, generating coherent and informative summaries from complex real-world inputs that integrate wearable data (e.g., simple statistics or

[†]These capabilities were investigated in a controlled, academic context. They do not indicate real-world products that require an evaluation of complex regulatory landscapes.

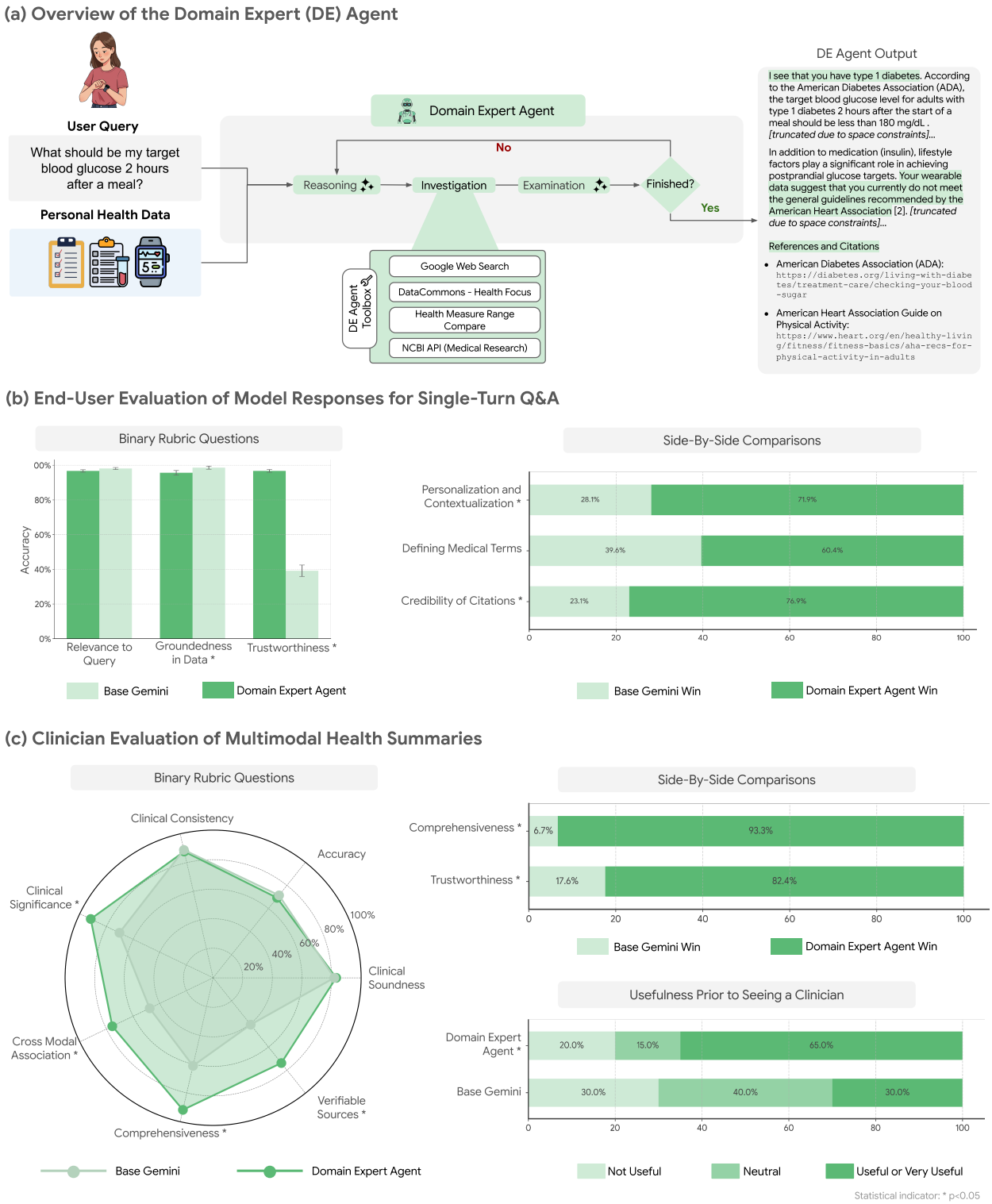


Figure 4 | Schematic and evaluation results of the *Domain Expert (DE) Agent*. **(a)** Overview of the DE Agent's workflow, which takes a user's query and personal health data as input. The agent employs an iterative process of reasoning, investigation, and examination, using a toolbox with access to specialized resources (see **Supplemental Section D.2** for prompting details). **(b)** End-user evaluation (N=17) of the DE Agent significantly outperforms a base Gemini model in multimodal dimensions in single-turn, contextualized Q&A. **(c)** Clinician evaluation (N=5) of multimodal health summaries shows same conclusions. **Supplemental Section D.3** summarizes evaluation details and rubrics.

complex outcomes computed by the DS agent) with medical records (e.g., lab results and medical history).

5.2. Domain Expert Agent Architecture

The DE Agent employs a multi-step planning, reasoning, and acting architecture (Yao et al., 2023), as shown in **Figure 4(a)**. The agent receives a user query, wearable data, and medical records as input. The agent processes the query and chooses actions from a set of tools that include querying: Web Search (Google, 2025), the National Center for Biotechnology Information (NCBI, 2025) application programming interface (API), and the DataCommons API (Data Commons, 2025). A Python Sandbox is also available for executing APIs.

The multi-step reasoning framework enables the agent to address complex queries by interleaving reasoning with tool use. For any given query, the agent follows a Reason-Investigate-Examine cycle: It begins with a reasoning and thinking stage, where the agent decomposes a user's health query into a series of verifiable steps, such as establishing a clinical baseline for a vital sign, finding literature to connect symptoms, or planning to analyze trends in personal data. Based on the reasoning from the previous step, the agent chooses the most suitable investigations by selecting the appropriate tool to execute its plan—for example, querying DataCommons for population statistics or NCBI for medical studies. The agent then examines the results of the previous steps, which informs the next cycle. This iterative cycle continues until the agent has synthesized sufficient evidence to construct a comprehensive answer.

This dynamic, iterative process creates a transparent and auditable reasoning trace. By synthesizing facts from authoritative sources with the user's personal health data, the agent ensures its final answer is not a black-box output but is instead a grounded, evidence-based conclusion. This architecture is therefore essential for fulfilling the agent's core capabilities of providing accurate, personalized, and reliable health insights. **Supplemental Section D.2** details the prompt setup for the DE Agent.

5.3. Domain Expert Agent Evaluation

We conducted a systematic evaluation process to evaluate our DE Agent's four capabilities in four independent tasks.

5.3.1. Task 1: Ground Truth of Knowledge Taxonomy and Factuality

Evaluation Goal. We evaluate our agent's ability to accurately answer nuanced medical and wellness questions (DE1). The evaluation consists of a comprehensive suite of expert-level multiple-choice questions covering personal health and medical knowledge domains.

Evaluation Setup. Guided by clinical relevance and our user study, we selected endocrinology, cardiology, fitness, and sleep medicine as the main domains to evaluate for this task. These domains map directly to real-world daily health needs, including both high-prevalence chronic conditions (endocrinology, cardiology) and high-interest preventative wellness topics (fitness, sleep). Furthermore, they provide a scientifically robust testing ground; their complexity requires deep reasoning over case vignettes, allowing us to assess performance beyond simple fact retrieval. While a broad medical knowledge base is essential, the agent's foundational language model has already demonstrated strong performance on general medical benchmarks (e.g., (Saab et al., 2024; Singhal et al., 2023). Our evaluation is therefore specifically designed to assess the *DE agent's* ability to apply this foundational knowledge in challenging, specialist-level domains that are of high interest to the public, rather than re-validating the base model's established capabilities.

To this end, using board certification and coaching exam questions provides a standardized and quantitative metric for evaluating the DE Agent's ability to use its medical knowledge to reason

over case studies mirroring real-world scenarios. Specifically, we took four curated multiple choice question (MCQ) datasets totalling more than 2000 test questions to assess knowledge and capabilities at answering expert-level health questions:

- **Endocrinology Examination:** Leveraging StatPearls' "American Board of Internal Medicine: Endocrinology, Diabetes, & Metabolism Exam" preparatory quizzes (StatPearls Publishing, 2024), we selected 570 questions at random from all available levels of difficulty (173 "Expert", 198 "Difficult", and 199 "Moderate and Easy" questions).
- **Cardiology Examination:** We created a list of 399 Cardiology Board Certification questions using BoardVital's ABIM-based preparatory question bank (BoardVitals, 2024). We randomly selected questions from all difficulty levels (100 "Hard", 199 "Moderate", and 100 "Easy" questions).
- **Fitness "Coaching" Certification Exam:** Compilation of 542 questions by Khasentino et al. (2025) from multiple question banks that emulate exam content for the Certified Strength and Conditioning Specialists.
- **Sleep Medicine Examination:** A curated dataset of 634 multiple choice questions from BoardVitals sleep medicine Maintenance of Certification board review question banks by Khasentino et al. (2025).

We compare our DE Agent against the same backbone model (Gemini 2.0 Pro), which acted as the baseline without the agentic framework.

Evaluation Results. The DE Agent's framework improves the base model's performance across all datasets and across almost all difficulty levels: the DE Agent achieves an overall accuracy of 83.6% while the base Gemini achieves 81.8%, showing a statistically significant improvement across all questions ($\chi^2 = 9.506$, $p = 0.002$, McNemar's Test, Odds Ratio 1.667). The detailed performance metrics and statistics are summarized in **Supplemental Table S10**.

5.3.2. Task 2: Diagnostic Conversations

Evaluation Goal. This task evaluates the agent's ability to perform differential diagnosis through conversational reasoning (DE2). Unlike prior work that often dealt with complex case studies (Kanjee et al., 2023; Liu et al., 2020d; McDuff et al., 2025), we focus on the agent's performance when presented with common symptoms that a user would typically self-report. The goal is to assess the quality and accuracy of the generated differential diagnosis based on these realistic scenarios.

Evaluation Setup. We conducted a large retrospective observational study to collect health case self reports from consumers (N=2,000, aged 18 - 65, IRB approved #GH-SCD-001). The collected data were de-identified and utilized to evaluate the performance of the DE Agent. Participants had experienced symptoms in one of 39 pre-defined categories (see **Supplemental Section D.4.2** for the full list) within the 3 months. Participants completed an online survey detailing their prior health events, their symptom descriptions in natural language, any online information-seeking behavior they engaged in, and the ultimate diagnosis they received.

We compared our DE Agent against a baseline that we re-implemented: a Gemini 2.0 Pro-based state-of-the-art DDx Agent (McDuff et al., 2025). To generate predictions, we input all self-reported symptom descriptions into each model and prompted for top-10 most likely diagnoses (refer to **Supplemental Section D.2.4** for the exact prompts). To evaluated the performance, we leveraged a separate Gemini 2.5 Pro as an autorater (prompt details in **Supplemental Section D.2.4**) that systematically compare the ground truth against the texts of predictions ranked by the models.

Evaluation Results. Our comparative analysis focuses on the top-10 predictions. The results showed that the DE Agent significantly outperforms the base model on the diagnostic capabilities ($p <$

0.001, $r = 0.156$, Wilcoxon signed-rank test, **Supplemental Table S11**). The DE Agent achieves a top-1 accuracy of 46.1% (whereas the DDx Agent achieves a top-1 accuracy of 41.4%, a performance advantage Δ=4.7%). In top-5 accuracy, the DE Agent reached 75.6% (Δ=3.8%), and its top-10 accuracy was 84.5% (Δ=3.1%, see **Supplemental Figure S6**). Our results showcase that the domain expert agent can accurately predict diagnoses from common self-reported symptoms, outperforming the state-of-the-art agent on differential diagnosis, underscoring its potential for enhanced utility in practical, user-facing health query and discussion scenarios.

5.3.3. Task 3: Contextualized and Personalized Responses to Health Queries

Evaluation Goal. Generic medical advice, while potentially accurate in isolation, can be inappropriate or even harmful if it fails to account for an individual's unique health profile. For instance, encouraging a vigorous exercise plan could be beneficial for most people but may pose significant risks to someone with certain cardiovascular conditions. Therefore, a critical capability of a reliable health agent is its ability to tailor responses to user-specific queries, integrating contextual information to provide personalized, trustworthy, and actionable guidance. This evaluation assesses the DE Agent's ability to personalize responses to health queries involving key contextual information (DE3).

Evaluation Setup. We curated a set of 50 health-related questions that are modified from the query set in Section 2.1 and designed to elicit personalized responses. To test the models' contextual reasoning, each query was augmented with specific personal information falling into two categories:

- **Demographic Differences:** Queries included details such as age, biological sex, and race/ethnicity, which critically influence health risks, disease presentation, and preventative care recommendations (e.g., screening schedules, risk stratification).
- **Existing Health Conditions:** Queries were framed within the context of prevalent chronic conditions, including diabetes, cardiovascular disease, and impaired kidney, thyroid, or liver function. Providing safe and relevant advice in these scenarios requires the agent to reason about potential contraindications and comorbidities.

We generated responses to these contextualized queries from both our DE Agent and the base Gemini 2.0 Pro model. To mitigate presentation bias, all responses were programmatically normalized to a standard format. A cohort of 17 end-users (aged 25 - 50) were recruited and evaluated the model responses in a randomized order where they were blind to the model condition. In total, end-users spent 13 hours on the evaluation. Inspired by Mallinar et al. (2025), our evaluations used *precise boolean rubric questions*, as well as side-by-side comparisons to assess dimensions of "personalization", "relevance", "credibility", and "trustworthiness". See **Supplemental Table S8** for the specific evaluation rubric.

Evaluation Results. Our evaluation revealed that while the base model could produce relevant information, the DE Agent was substantially more effective at generating more personalized, trustworthy, and well-supported responses and guidance essential for health applications, as shown in **Figure 4(b)**.

In binary rubric assessments, end-users had good to excellent IRR (Fleiss' $\kappa \geq 0.684$). DE Agent's received significantly higher Trustworthiness ratings (96.9±0.8%), whereas the base Gemini model's responses achieved only 38.7±3.3% ($p < 0.001, \beta = 6.873$, Binomial generalized linear mixed model, GLMM). Menawhile, both models performed similarly on measures of Relevance-to-Query rubric (97.6±0.7% and 98.1±0.6%, respectively, $p = 0.465, \beta = -0.269$) and Groundedness-in-Data, 95.6±1.3% and 98.6±0.9, $p < 0.001, \beta = -1.605$), indicating that both models successfully identified the core intent of the user's question and provided solid data-driven responses. **Supplemental Table S12** presents more statistical details. These results highlight a critical failure of the general-purpose model to generate content that users perceive as reliable for health decisions.

Side-by-side comparisons further highlighted the DE Agent's ability to personalize responses compared to the base model. When asked to choose the better response, end-users preferred the DE Agent over the base model in 71.9±1.5% of cases in the Personalization-and-Contextualization dimension ($p < 0.001, \beta = 1.667$). Furthermore, the DE Agent also showed better performance at Defining-Medical-Terms, winning 60.4±1.7% of comparisons ($p = 0.057, \beta = 0.664$), and won 76.9±1.5% ($p < 0.001, \beta = 1.969$) of comparisons in Credibility-of-Citations rubric. See **Supplemental Table S13** for more detailed results.

These outcomes showed that the DE Agent's design and domain-specific tools enable it to synthesize personal context with verifiable medical knowledge to produce more nuanced, comprehensible, and trusted responses compared to the base model.

5.3.4. Task 4: Reasoning over Multi-Modal Personal Health Data

Evaluation Goal. This task evaluates the agent's ability to synthesize and reason over diverse, multimodal personal health data (wearable data and medical records) to generate comprehensive health summaries (DE4). Unlike simple information retrieval, this task demands a higher level of cognitive synthesis that mimics a clinician's ability to connect disparate data points, identify trends, and highlight potential concerns. The objective is to produce a concise and actionable overview of an individual's health status proactively, without a specific user query or chief complaint.

Evaluation Setup. To evaluate our agent's capabilities for this task, we leverage the following data modalities from participants from the WEAR-ME data:

- **Unstructured Text:** For each participant, we combine all contextual and background information collected during onboarding survey as one string.
- **Structured Tables:** This included tabular data representing wearable-derived digital measures, such as resting heart rate, heart rate variability, etc.
- **Lab Result Documents (PDFs):** Given that most users will have access to an electronic document of their lab tests, we convert real participants' lab results to PDF documents for input to the compared models.

We sampled 30 participants from our persona pools defined in Section 3.4 with an equal distribution across 10 health profiles. We then prompted our DE Agent and the baseline Gemini 2.0 Pro model to generate a comprehensive health summary for each participant based on their complete data profile. A panel of 5 expert clinicians (MDs, 2 females, 3 males, aged 30 - 60) then independently evaluated the quality of these generated summaries in a side-by-side, blinded comparison, with a total of 77 hours of human effort. The clinicians used a detailed rubric to assess dimensions such as clinical relevance, accuracy of interpretation, and the actionability of the summary (see **Supplemental Table S9** for the full rubric).

Evaluation Results. Our results demonstrate that the DE Agent generates comprehensive, clinically relevant, and useful multimodal health summaries compared to base model. As shown in **Figure 4(c)**, on binary rubric questions, the DE Agent showed significant improvement in Clinical-Significance (96.4±0.6% vs. 73.8±0.6%, $p < 0.001, \beta = 2.461$, Binomial GLMM), Cross-Modal-Association (79.4±16.9% vs. 50.0±15.0%, $p < 0.001, \beta = 3.355$), Comprehensiveness (96.4±1.1% vs. 64.2±11.8%, $p < 0.001, \beta = 3.024$), and Citation-Credibility (77.6±13.8% vs. 42.7±17.9%, $p < 0.001, \beta = 3.721$). More statistical details are listed in **Supplemental Table S14**. In direct side-by-side comparisons, clinicians overwhelmingly preferred the summaries generated by the DE Agent over those from the base model. When asked for an overall preference on Comprehensiveness, 93.3±1.9% of evaluations favored the DE Agent over the base model ($p < 0.001, \beta = 2.652$). On Trustworthiness, the DE Agent was again strongly preferred, winning 82.4±3.0% of comparisons ($p < 0.001, \beta = 3.721$, **Supplemental Table S15**). These results underscore a clear preference for the

DE Agent's output in a holistic assessment.

Another significant differentiator also emerged in the clinical usefulness of the summaries prior to seeing a clinician. As shown in **Figure 4(c)**, the DE Agent was rated as substantially more useful ($p < 0.001, \beta = 0.252$, **Supplemental Table S16**): Approximately, 65.0% of its summaries were rated as "Very Useful" (5 on a 5-point scale) or "Useful" (4 on the scale). In contrast, only 30.0% of base model's summaries achieved these top ratings. Conversely, base model's summaries were more often rated at the lower end of the scale, with 30.0% being rated "Not Useful", compared to only 20.0% for the DE Agent. In addition, the safety profile of the DE Agent's summaries was also highly favorable by the experts (**Supplemental Figure S7 and Figure S8**).

Collectively, these expert evaluations demonstrate that DE Agent not only synthesizes multimodal health data effectively but also produces summaries that are perceived by clinicians as more clinically significant, comprehensive, safe, and significantly more useful for patients than those generated by current leading general-purpose multimodal models.

6. Health Coach Agent

A personal health agent needs to be equipped with robust coaching capabilities to address user needs for goal setting and wellness planning. Early research on conversational agents predominantly focused on single-turn, factoid question-answering, where response quality could be assessed objectively by third-party experts (Bai et al., 2024; Deng et al., 2024). While recent advances have extended their application to health applications, such as analyzing conversations for mental health and counseling (Althoff et al., 2016), evaluating different types of therapies and interventions (Mehta et al., 2022), and facilitating cognitive reframing of negative thoughts (Sharma et al., 2023, 2024), the specific requirements of personalized health coaching remain a distinct challenge.

Effective health coaching presents several unique challenges not fully addressed by traditional conversational models (Deng et al., 2024). First, user goals are often open-ended and evolve during the interaction. Second, the initial context provided by the user is typically ambiguous and incomplete, necessitating iterative clarification over multiple turns. Third, coaching requires a mixed-initiative dialogue where the agent needs to balance user-led exploration with proactive, coach-led guidance to gather information about goals, preferences, and barriers. These complexities are compounded by variations in user conversational styles and the potential for off-topic diversions.

Beyond these structural challenges, the perceived quality of a coaching interaction is highly dependent on user engagement and satisfaction. A critical aspect of this is managing the timing of recommendations. Premature recommendations without sufficient information gathering can seem dismissive and impact user trust. Conversely, excessive probing without offering actionable advice can frustrate the user and diminish their confidence in the agent's capabilities (Ikemoto et al., 2018). Therefore, our **Health Coach (HC) Agent** is explicitly designed to navigate this trade-off. By adopting a structured, expert-inspired approach to multi-turn conversation, our goal is to enhance user engagement and satisfaction, thereby providing a robust coaching component for our multi-agent system.

6.1. Health Coach Agent Capabilities

To inform the design of our HC Agent, we conducted a formative interview study with 11 health experts (6 males, 5 females, aged 24 - 45). Experts were selected from various domains such as sleep, fitness and life coaching. These experts hold advanced degrees in their respective fields (with at least 5 years of experience) and had a deep understanding of best practices in health coaching. The semi-structured interviews, each lasting approximately one hour, focused on eliciting expert insights regarding the critical elements of effective health coaching, particularly focusing on the interplay of conversational substances and styles. Two authors conducted thematic analysis and

independently developed theme codebooks and iterated until convergence. This qualitative data provided a foundational understanding of nuanced coaching expertise, leading to the following core HC Agent capabilities:

HC1. **Goal and purpose identification**: The HC Agent needs to understand users' goal and the motivation behind the goal, maintain a goal-oriented conversation, and bring the user back to the main goal when they are deviating off the topic.

HC2. **Active listening**: The agent should provide occasional summary and paraphrasing of the state of conversation to circle back to users and ensure correct understanding and goal alignment.

HC3. **Context clarification**: Our HC Agent should collect and understand users' constraints, preferences, and previously attempted solutions.

HC4. **User empowerment**: The HC Agent needs to empower the user by building trust, guiding them to navigate their own paths when exploring potential solutions, and providing supportive reinforcement when they make positive progress toward their goals.

HC5. **Personalized and concrete coaching recommendations**: The HC Agent should provide high-quality customized recommendations that are specific, measurable, attainable, relevant, and time-bound (SMART), helping the user create an actionable plan (Doran, 1981).

HC6. **User feedback incorporation**: The agent should be able to provide initial recommendations mid-conversation, while explicitly leaving space for user input, and subsequently updating its suggestions based on that feedback.

6.2. Health Coach Agent Architecture

Conventional single-prompt architectures often fail to simultaneously address the diverse coaching capabilities (HC1-HC6), struggling in particular with the demands of mixed-initiative dialogue. To overcome these limitations, we designed a modular architecture for the HC Agent that embodies expert coaching principles. As illustrated in **Figure 5(a)**, this design enables the agent to dynamically adapt its conversational flow based on user input and to strategically time its recommendations.

Personalized Coaching Module. At its core, the HC Agent is designed to emulate a professional coach. To this end, we developed the *Personalized Coaching Module* grounded in the principles inspired by motivational interviewing (Hettema et al., 2005), a client-centered approach that enhances motivation by exploring and resolving ambivalence. This module operationalizes key coaching behaviors, such as active listening (HC2) and user empowerment (HC4), to foster engagement. It also establishes a structured, proactive line of questioning that instructs the agent to first identify the user's goals (HC1), then explore their context and constraints (HC3), and finally solicit feedback (HC6) before providing a recommendation (HC5).

Conversation Flow Modules: Recommendation Module & Conclusion Module. While a single prompt can include all these aspects, our pilot testing revealed that such a prompt-based approach struggles to maintain conversational balance. Such agents tended to either offer premature recommendations before sufficient context was gathered, or become fixated on questioning, failing to transition to actionable advice. To address these failure cases, we introduce two specialized modules to manage the conversational state:

The *Recommendation Module* dynamically determines the opportune moment to provide advice. It is responsible for keeping the conversation on track (HC1) and explicitly confirms with the user if

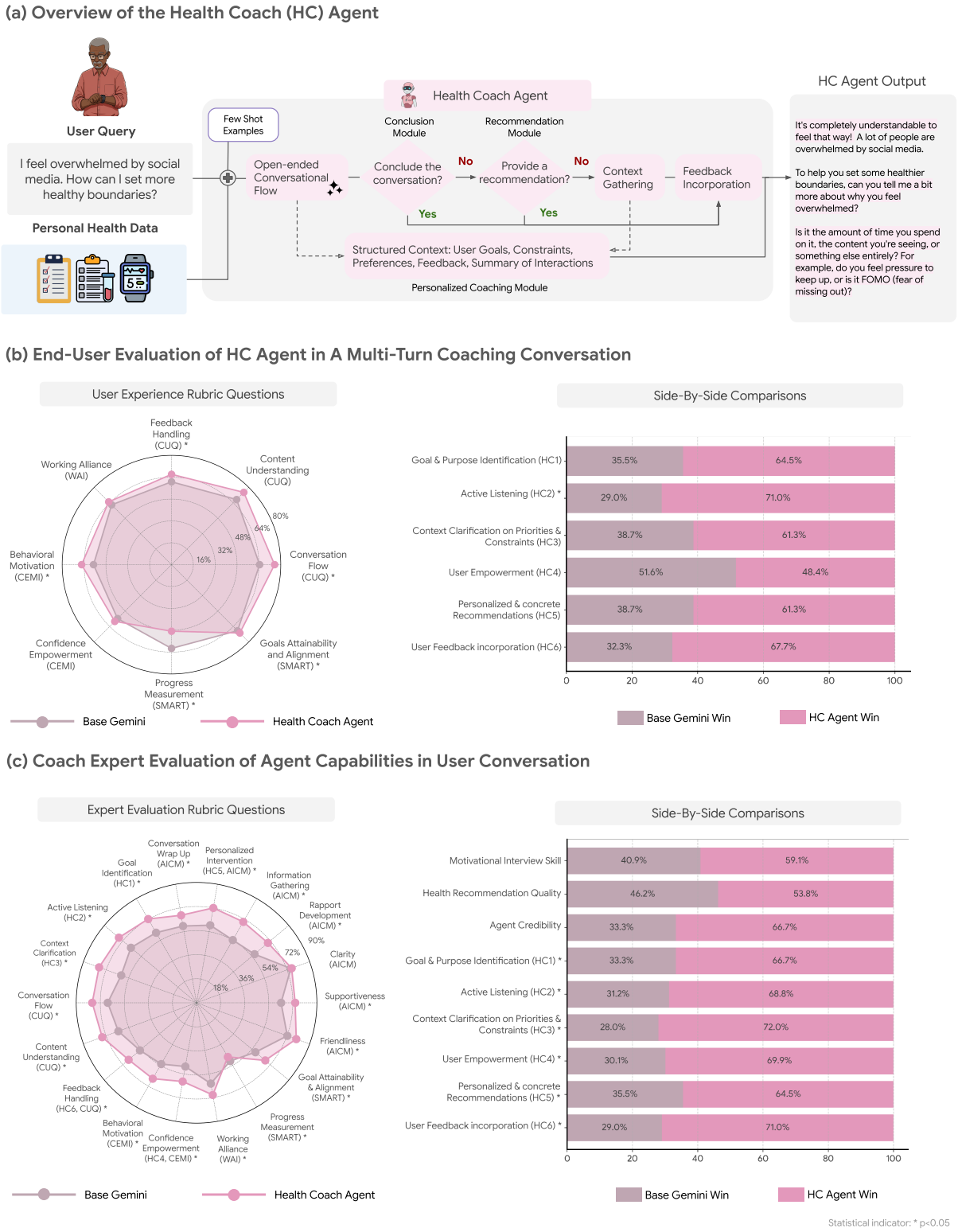


Figure 5 | **Schematic and evaluation results of the *Health Coach (HC) Agent*.** **(a)** The HC Agent architecture employs a modular process with a core Personalized Coaching Module and Conversation Flow Modules (Recommendation and Conclusion) to adapt the dialogue and strategically time recommendations (Supplemental Section E.2 shows prompting details). **(b)** End-user evaluation (N=31) shows the HC Agent significantly outperforms a base model in multi-turn coaching conversations, leading to higher user engagement and preference. **(c)** Coach expert evaluation (N=3) of conversation transcripts confirms these conclusions, revealing an even more pronounced advantage for the HC Agent from experts' perspectives (more details in Supplemental Section E.3).

enough context has been gathered before transitioning to a recommendation (HC3, HC6). Meanwhile, the *Conclusion Module* identifies when a conversation has reached a natural endpoint and ends the interaction gracefully with a summary, reinforcing the key takeaways (HC5).

The final architecture of the HC Agent is a combination of the Personalized Coaching Module to steer the coaching process with personalized information, and the two conversation flow modules for appropriate recommendations and conclusion to align the flow of the interaction with a user's queries, inputs, and goals. More prompting details are described in **Supplemental Section E.2**. Overall, this modular architecture ensures the agent robustly expresses the full range of expert coaching themes (HC1-HC6) in a dynamic, user-adaptive manner.

6.3. Health Coach Agent Evaluation

To assess the HC Agent's capabilities, we conducted a comprehensive evaluation centered on real-world user interactions across a variety of health and wellness topics. For a holistic assessment, our evaluation incorporates perspectives from both end-users and coaching experts.

6.3.1. Task 1: Health Coach Agent Capability from End-Users' Perspectives

Evaluation Goal. The practical success of an LLM-based health coach hinges on the end-user's experience. An agent is effective only if users perceive it as engaging, competent, and preferable to alternatives. This evaluation, therefore, investigates user sentiment and perception of the HC Agent versus a baseline. We measure overall preference and engagement via agent rankings, satisfaction scores, and conversation outcomes (i.e., natural vs. forced endings). To understand the drivers of these preferences, we also assess user perceptions of specific coaching capabilities, conversational style, and the actionability of the agent's guidance.

Evaluation Setup. We recruited 31 participants (25 males, 6 females, aged 18-64) for a 1.5-hour session each (48 hours of human effort total). Participants interacted with multiple conversational agents in a blinded, counterbalanced order using a Latin Square design. This paper focuses on the comparison between our HC Agent and a baseline agent built on the same Gemini model, prompted with a general instruction to act as a helpful personal coach. No significant difference in response time was observed between agent conditions. A full comparison of all agent designs is available in Srinivas et al. (2025).

To guide the interactions, we curated a set of open-ended scenarios reflecting common health concerns (see **Supplemental Table S17**). Importantly, participants selected the scenario that best aligned with their own personal interests to ensure a realistic coaching use case. They were instructed to treat each agent interaction as a new consultation, providing consistent background information each time (see **Supplemental Section E.3.2** for detailed evaluation instructions). Each interaction began with the same initial sentence from the chosen scenario, after which participants conversed freely until reaching either a natural resolution or a forced ending (if the conversation was deemed unproductive).

Participants assessed each agent post-interaction and provided a final ranking at the end of the study. We adopted established metrics from psychology and agent usability, including subsets of the Chatbot Usability Questionnaire (CUQ, Holmes et al. (2019)), the Working Alliance Inventory (WAI, Munder et al. (2010)), the Client Evaluation of Motivational Interviewing (CEMI, Madson et al. (2013)), and SMART goal setting evaluations (Doran, 1981). These metrics provide a comprehensive view of usability, collaborative bond, motivational techniques, and goal formulation, echoing the expert-identified coaching themes (see **Supplemental Table S18 for raw questions and Table S20** for detailed aspects of these metrics). By allowing user choice in scenarios and enabling open-ended interactions, this setup is designed to capture realistic user experiences that surpass the limitations of

automated evaluations.

Evaluation Results. Overall, participants rated our HC Agent higher across most dimensions of the CUQ, WAI, CEMI, and SMART metrics compared to the baseline agent. As indicated in **Figure 5(b)**, the HC Agent exhibited better capability in controlling conversation flow (scored normalized as percentage for consistency: 64.5 ± 3.4% vs. 75.5 ± 3.3%, $p = 0.025, \beta = 0.164$, GLMM with Gamma link function), conducting motivational interviews (57.1 ± 3.8% vs. 65.8 ± 3.2%, $p = 0.005, \beta = 0.172$), and incorporating user feedback (60.7 ± 3.8% vs. 66.1 ± 2.9%, $p < 0.001, \beta = 0.111$). Full statistical details can be found in see **Supplemental Table S24**.

The advantage of our HC Agent further reflected its higher rankings compared to the baseline agent. In five out of the six themes, participants preferred our agent over the baseline (win rates 61.3-71.0%, see **Supplemental Table S25** with detailed results). In particular, our agent showed significant advantage in Active-Listening (HC2, 71.0% win rate, $p = 0.025, \beta = 0.894$). Additionally, we found that the HC Agent resulted in 83.9% natural endings, while the baseline agent naturally concluded only 74.2% of interactions. This demonstrates that the HC Agent was able to promote higher user engagement. On the other hand, the baseline agent outperformed our proposed agent in the progress measurement aspect of the expert rubrics (61.3 ± 4.5% vs. 48.7 ± 3.9%, $p = 0.010, \beta = -0.230$) and had a similar ranking result in the user empowerment (HC4), suggesting further direction to improve in our final multi-agent system.

6.3.2. Task 2: Validation of Health Coach Agent Design with Expert Evaluation

Evaluation Goal. While end-users are well-equipped to judge their personal satisfaction and engagement (Task 1), they may not be able to recognize whether an agent is correctly applying evidence-based coaching techniques. A comprehensive evaluation should also validate the agent's methodological rigor. Therefore, an expert-led validation is essential to assess the fidelity of the HC Agent's design. The primary goal of this task is to verify that the agent's conversational behaviors are not just subjectively engaging but also objectively aligned with the intended expert-defined coaching capabilities (HC1-HC6).

Evaluation Setup. We recruited three of the health experts from our formative study (**Section 6.1**) to evaluate the conversation transcripts collected from Task 1. The experts, blind to the agent conditions, rated each conversation according to the six coaching themes (HC1-HC6), as well as the same set of CUQ, WAI, CEMI, and SMART questions. We also incorporated a specialized set of AI Coaching Metrics (AICM) developed by our health coach experts, covering conversation tones (supportiveness, friendliness), style (clarification), and strategy (rapport development, information gathering for key assessment, personalized intervention, appropriate summary, see **Supplemental Section E.3.3** for detailed evaluation instructions and questions). Finally, experts provided comparative rankings of the agents on motivational interviewing skills, recommendation quality, and credibility.

Evaluation Results. As shown in **Figure 5(c)**, our HC Agent outperformed the base model in almost all dimensions across expert themes ($\Delta = 11.0 - 12.8\%$, $ps < 0.001$, $\beta = 0.197 - 0.307$, GLMMs with Gamma link function), motivational interview ($\Delta = 11.3 - 12.8\%$, $ps < 0.001$, $\beta = 0.227 - 0.246$), AI coaching ($\Delta = 7.1 - 15.7\%$, $ps < 0.001$, $\beta = 0.106 - 0.302$, except that the two agents had a tie on Clarity, $p = 0.678$), Working-Alliance ($\Delta = 8.5\%$, $p < 0.001$), and Goal-Alignment ($\Delta = 9.4\%$, $p < 0.001$, $\beta = 0.192$). Compared to the end-user ratings (relative advantage = 3.5 ± 2.5%), the advantage of HC Agent is more apparent according to coaching experts (relative advantage = 10.2 ± 1.2%). The ranking results also suggests that the experts strongly preferred the HC Agent over the baseline according to the metrics on Motivational-Interview-Skills (win rate 59.1%), Recommendation-Quality (win rate 53.8%), Credibility (win rate 66.7%), as well as the themes (HC1-HC6, win rates 64.5-72.0%, $ps < 0.05$, $\beta = 0.649 - 1.190$). **Supplemental Table S26 and Table S27** lists statistical details

in all these metrics.

Consistent with end-users, experts also gave higher ratings to the baseline agent on Progress-Measurement ($50.3 \pm 2.4\%$ vs. $46.7 \pm 1.4\%$, $p < 0.001$, $\beta = -0.012$). This may suggest that, within a limited number of interactions in a conversation, the HC Agent's current design prioritizes and allocates more content to context gathering and goal alignment over progress tracking. In contrast, the base agent tends to comprehensively cover all aspects, although it underperforms the HC agent on most other aspects. In the next section, we refined the HC part within the multi-agent system to put more emphasis on progress tracking (see **Section 7**). Additionally, a notable discrepancy emerged in the User-Empowerment dimension: while end-users ranked the agents similarly (HC Agent win rate of 48.4%), experts strongly preferred the HC Agent (69.9% win rate). This highlights the multifaceted nature of coaching evaluation and underscores the importance of incorporating both expert and end-user perspectives.

Overall, our dual-perspective evaluation, incorporating both end-user and expert assessments, consistently demonstrates that our modular architecture produces an HC Agent with significantly stronger and more effective health coaching capabilities.

7. Personal Health Agent: A Multi-Agent Framework

While each agent possesses a unique and valuable capability (Section 4-Section 6), addressing the multifaceted nature of personal health requires their synergistic collaboration. To provide comprehensive support for users' diverse health needs and emulate the structure of real-world human health teams, we assemble the three specialized agents into a cohesive multi-agent system: the **Personal Health Agent (PHA)**. Any real-world application inspired by this research would require a separate and rigorous analysis of its features and claims to ensure compliance with all applicable regulations. Building on recent advances in multi-agent systems (Hong et al., 2024; Li et al., 2023; Wu et al., 2023), we propose a set of design principles for personal health agents, which adopt both the theoretical framework of modular cognition (Fodor, 1983) and empirical knowledge about human collaborative teams in the real world. The architecture we designed for our PHA is a direct embodiment of these principles. The assessment of our PHA completes the last piece of our comprehensive evaluation framework of a general personal health agent.

7.1. Multi-Agent Design Principles

Recent work has developed several multi-agent architectures and strategies that vary in complexity and are often inspired by different human decision-making approaches. Examples include voting (Wang et al., 2022) and debating (Du et al., 2024). Moreover, studies have shown that iterative multi-agent processes can improve the accuracy and robustness of agent outputs (Chen et al., 2024; Tang et al., 2024). Palepu et al. (2025) leveraged the dual-system theory to design a two-agent framework to guide disease management conversations. Inspired by human team formation, Kim et al. (2024) created a dynamic strategy to assign different sizes of a multi-agent team to balance computation cost and decision accuracy. These studies highlight the value of multi-agent collaboration, yet often focus on generic interaction patterns or homogenous agents. Inspired by the human-team approach, our work explores a functional design where each agent embodies a distinct professional role with unique capabilities, enabling a more structured and comprehensive approach to personal health inquiries.

Our design is grounded in both cognitive theory and empirical evidence from human collaboration. Theoretically, we draw from the concept of modular cognition, which posits that the mind comprises specialized, independent subsystems for different cognitive functions (Fodor, 1983). We apply this by treating each agent as a distinct cognitive module with its own knowledge and processing capabilities. Empirically, this mirrors the structure of effective human expert teams, where members

with complementary expertise collaborate to achieve a more robust outcome than any individual could alone. This modular approach not only enables a more holistic response to complex health questions but also enhances explainability. By tracing outputs back to the specific function of each agent, the system's decision-making becomes more transparent and interpretable, an essential requirement in sensitive health contexts.

To further instantiate our design, we propose a set of four principles across capability, user experience, and architecture to guide the design of our PHA:

P1. **Address Comprehensive User Needs**: Our user-centered analysis in Section 2 has revealed the wide range of users' health and wellness needs. This calls for a comprehensive set of agent capabilities to fulfill these needs. Single-agent systems, even with rich capabilities (e.g., Merrill et al. (2024a)), may not be able to cover all cases.

P2. **Adaptive Support**: We previously identified that user queries may require different combinations of the three agents (Supplemental Table S1). A multi-agent system needs to dynamically assign agents to be adaptive toward users' needs. This suggests that static multi-agent architectures, such as voting (Wang et al., 2022) and debating (Du et al., 2024), may not fulfill the need.

P3. **Low User Burden**: When a system has access to users' wearable data and medical records, asking users what can already be inferred from the data would not only increase user burden, but also impact user trust of the system's intelligence. For example, the system should avoid asking questions such as "how long did you sleep last week?" (a typical question from the HC Agent when it acts without the DS Agent) when it can easily compute the data from the wearable data record (via the DS Agent). The system needs to have a chain of agents for deeper reflection rather than simply a set of parallel agents for straight response, e.g., voting (Chen et al., 2024; Wang et al., 2023), to fully leverage and potentially follow up on the output from each agent.

P4. **Simplicity**: When agent's capabilities are similar, simpler architecture has advantages of lower latency, less computation cost, as well as easier pathway of maintenance and extension (Schaffer, 2015). While more complex, hierarchical structures can be effective for specific goals like ensuring clinical safety (Kim et al., 2025b), or mimicking interdisciplinary care teams (Kim et al., 2024), they may represent unnecessary complexity (e.g., leading to unappealing long inference times).

These four guidelines indicate the limitations of existing multi-agent architectures and guide the implementation of the multi-agent collaboration framework of PHA.

7.2. Multi-Agent Collaboration Framework

Following theoretical foundations, empirical evidence, together with the design principles, **Figure 6(a)** presents our multi-agent framework of PHA. It consists of an *Orchestrator* (Gemini-based) that manages a sequential and iterative workflow between the three specialists agents: **the DS Agent** for data science analysis, **the DE Agent** for domain expertise, and **the HC Agent** for health coaching (improved on progress tracking based on the results in **Section 6**). This collaborative process involves four key steps. The implementation details can be found in **Supplemental Section F.2**:

1. **User Need Understanding (P1, P2):** When a user query arrives, the Orchestrator first performs the crucial step of user need understanding. It analyzes the semantics of the query and infers the user's primary goal. This classification aligns with the four CUJs identified in Section 2 (e.g., seeking data insights, wellness advice). This ensures that the system can select the most

appropriate agent(s) for the user's need, addressing the principles of covering Comprehensive User Needs (P1) and providing Adaptive Support (P2).

2. **Agent Orchestration and Task Assignment (P2, P4):** Based on the inferred user need(s), the Orchestrator dynamically assigns a "main agent" (one of the three agents) to take the primary responsibility for the query. It then determines if any "supporting agents" are required to provide supplementary information. For instance, a query like, "How can I improve my sleep based on last week's data?" would assign the HC Agent as the main agent, supported by the DS agent to first analyze the sleep data. The Orchestrator decomposes the main query into specific, targeted sub-queries for each assigned agent. Moreover, to facilitate the task assignment, we built upon the user journey categories in Section 2.2 and developed a set of collaboration examples that denote agent assignment for each sub-category (see Supplemental Table S1). These examples are provided to the Orchestrator for reference. This dynamic needs-based assignment fulfills the principle of Adaptive Support (P2) while maintaining architectural Simplicity (P4) by avoiding a complex hierarchy.
3. **Query Reflection (P3):** Unlike parallel architectures where agents work in isolation, PHA employs an iterative process that allows for reflection. Outputs from the main agent and the supporting agents are not sent directly to the user. Before finalizing the response, the main agent performs a "self-query reflection" step. Inspired by iterative self-refinement frameworks (Chen et al., 2024; Kim et al., 2025a; Madaan et al., 2023), this internal check assesses the synthesized information for coherence, completeness, and accuracy, ensuring all available data from the user's health data is leveraged. By preventing the system from asking for data it already has (e.g., "How many hours did you sleep? / When did you wake up?"), this process upholds the principle of Low User Burden (P3).
4. **Memory Update (P1, P2, P3):** After each interaction turn is completed and a response is presented to the user, the Orchestrator updates a persistent memory module. This memory logs key entities from the conversation, including generated insights from each agent, user-stated goals, identified barriers, and preferences. This running context enables the system to maintain conversational coherence over multiple turns and personalize future interactions more effectively. The memory module supports the agent team in addressing Comprehensive Needs (P1) over time, providing Adaptive Support (P2) as goals evolve, and ensuring Low User Burden (P3) by retaining context across sessions.

Our proposed PHA represents a concrete and principled step toward the broader vision of integrated, end-to-end personal health agents. While our evaluation in the next section demonstrates its efficacy against current baselines, we present this architecture not as a definitive solution, but as a foundational blueprint. We anticipate this work will catalyze future research into alternative collaborative strategies and more advanced agent designs, building upon the principles and findings established in this work.

7.3. Personal Health Agent Evaluation

For a comprehensive evaluation, we compare our PHA against two baseline agent setups (with the same Gemini 2.0 Pro backbone) to demonstrate the value of our collaborative architecture.

- **Baseline 1: Single-Agent Personal Health Insights Agent (PHIA) System.** Prior work has revealed the promising capability of a single agent to conduct data analysis, generate personal health insights, and provide health recommendation. We re-implemented the architecture in Merrill et al. (2024a) for a PHIA system that is capable of code generation, information retrieval, online search to analyze and interpret behavioral personal health data. This system covers the key capabilities we have identified for the three specialist agents. Note that although this is a single-agent system, its ReAct framework can still trigger multiple LLM calls (Yao et al., 2023). While this design embodies architectural simplicity (+**P4**), it may struggle to cohesively

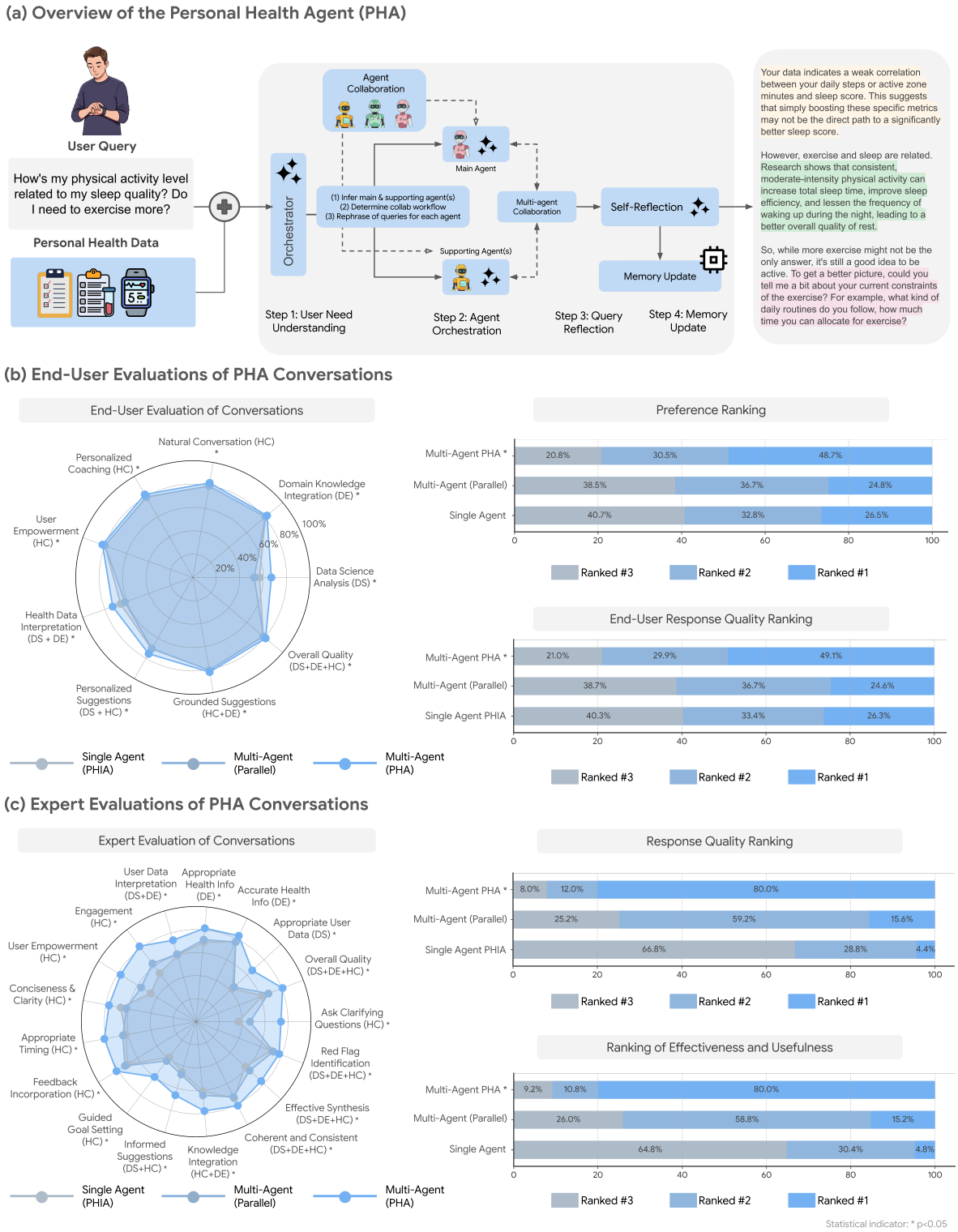


Figure 6 | Schematic and evaluation results of the *Personal Health Agent (PHA)*. **(a)** Overview of the PHA's collaborative workflow. An Orchestrator dynamically assigns tasks to the DS, DE, and HC agents, managing a multi-step process of collaboration, reflection, and memory update to synthesize a cohesive response (see Supplemental Section F.2 for prompting details). **(b)** End-user evaluation (N=20) of conversations across 50 user personas from WEAR-ME shows the PHA is significantly preferred over both single-agent and parallel multi-agent baselines for overall quality and helpfulness. **(c)** Health expert evaluation (N=5) of the conversation transcripts enhance these conclusions, revealing a substantially more pronounced preference for the PHA system over the baselines. Supplemental Section F.3 summarizes more evaluation details.

integrate the distinct skills of data science, domain expertise, and health coaching within a single agent (−**P1**, −**P2**, −**P3**).

- **Baseline 2: Parallel Multi-Agent System** For multi-agent baselines, we implemented a static architecture that adopts a voting-like strategy (Chen et al., 2024; Wang et al., 2023): each of the three agents will respond to a user's query in parallel, and an orchestrator will combine the responses and respond to the query. This setup leverages the capabilities of each agent (+**P1**) and ensures a short response time with a simpler structure than PHA (+**P4**). However, different agents are working independently with a static structure, which may not be flexible enough to handle diverse users' needs (−**P2**). It can miss the opportunity to fully unlock the synergies among three agents and lead to a less useful response (−**P3**).

7.3.1. Task 1: Multi-agent Conversation Experience for End-Users

Evaluation Goal. After combining the three single agents into a multi-agent system, our first evaluation is to assess PHA's ability to conduct broad, multi-turn health-related conversations from an end-user's perspective. We aim to measure the perceived quality of the interaction, focusing on metrics such as helpfulness, coherence, empathy, and how well the agent integrates data analysis, domain knowledge, and health coaching.

Evaluation Setup. We adopted a controlled, role-playing methodology to mitigate the potential influence introduced by varied agent response time, user backgrounds, and conversation directions, ensuring the consistency across evaluations.

Conversation Generation: As introduced in Section 3.4, we sampled a set of 50 representative user personas from the WEAR-ME dataset, covering 10 common health profiles (5 personas each), ranging from healthy individuals training for a goal to individuals managing chronic conditions like hypertension or diabetes (examples in Supplemental Section B.2.1). Working with primary care physicians, we developed detailed background stories and specific health needs for each persona. We then used a base Gemini 2.0 Flash model, instructed with a persona's data, background and goals, to act as the user and engage in three independent, multi-turn conversations with each of the three systems: the single-agent baseline, the parallel multi-agent baseline, and our PHA. All three systems had access to the same persona data. Conversations concluded when the role-playing user agent determined its needs were met. This results in a total of 150 conversation scripts (50 personas × 3 agent systems).

Conversation Evaluation: We then recruited 20 end-user generalist evaluators (13 males, 7 females, aged 20 - 44) with prior experience on using daily wearable devices and reading data related to personal health. For each of the 50 personas, an evaluator was first asked to read the user's data, background story, and goals to put themselves in the shoes of this individual. They then read the three corresponding conversation scripts (raters were blind to agent conditions), presented in a randomized order to mitigate primary and recency effects. After reading each script, evaluators completed an evaluation form. **Supplemental Section F.3.1** contains the detailed instruction. The metrics include the agent's overall quality and its specific capabilities in data science, domain knowledge, and coaching, as well as the organic integration of these skills (see specific questions in **Supplemental Table S28** and rubrics in **Supplemental Table S29**). After rating all three agents for a single persona, evaluators provided a ranking of the three agents. This process resulted in 3,000 individual agent ratings (150 scripts × 20 evaluators) and 1,000 comparative rankings (50 personas × 20 evaluators). In total, 20 end-users spent 500 hours on the evaluation.

Evaluation Results. **Figure 6(b)** summarizes the results from our end-user evaluation. The findings indicate that while all three systems were perceived as generally competent, our proposed PHA architecture was consistently favored for providing a better overall experience.

When asked to select the best overall conversation, users ranked PHA as the top nearly half the time

(48.7% based on the overall preference, and 49.1% based on the response quality), significantly outperforming the parallel (24.8% and 24.6%) and single-agent (26.5% and 26.3%) baselines ($ps < 0.001$, Friedman tests followed by post-hoc Wilcoxon Rank-sum tests with Holm-Bonferroni correction). The detailed breakdown of agents' performance in the radar chart illuminates more detailed insights about PHA's strengths. PHA outperform both baseline agents in all dimensions with statistical significance ($ps < 0.001$). In particular, PHA has a clear advantage for Personal-Data-Analysis (70.3 ± 0.7% vs 54.9 ± 0.8% vs 59.7 ± 0.8% for PHA, parallel, and single agent, $p < 0.001$, $\beta_{PHA-PHIA} = 0.175$, $\beta_{PHA-Parallel} = 0.255$, GLMM with Gamma link function and post-doc Tukey's HSD) and the Data-Results-Interpretation (the integration of DS and DE Agents, 76.5 ± 0.6% vs 64.8 ± 0.8% vs 69.1 ± 0.8%, $p < 0.001$, $\beta_{PHA-PHIA} = 0.114$, $\beta_{PHA-Parallel} = 0.182$). Detailed results can be found in **Supplemental Table S32 and Table S33**.

This illustrates that the primary value of PHA, from an end user's perspective, lies in its unique ability to bridge the gap between quantitative data analysis and qualitative health interpretation. While the baseline agents may possess the skills on domain knowledge and coaching, PHA's iterative framework allows the DE agent to reflect upon and contextualize the specific findings from the DS agent, providing a more deeply synthesized and insightful response. This process is critical for translating complex personal health data into clear, meaningful, and actionable information for the end-user, directly addressing a core need identified in our formative research. Moreover, the clear advantage on the ranking further suggests that users may find individual responses from the baselines acceptable, the cumulative experience of a coherent, evolving conversation – a direct benefit of PHA's iterative reflection and memory – is a powerful differentiator that users clearly prefer. This is also reflected from the qualitative comments from evaluators. For instance, one rater stated, *"The conversation with [PHA] is most helpful because it provided the most specific, actionable, and tailored advice for both running and cholesterol goals, directly addressing stated preferences and concerns."* Another rater commented, *"[PHA] was the best because it immediately understood the user's biggest problem... and gave easy, personalized solutions using their own activity tracking information. [The two baselines] gave good advice too, but they didn't focus on that specific problem."* These results indicate the clear advantage of our multi-agent PHA system.

Interestingly, the single agent PHIA baseline, from an end-user perspective, has similar performance compared to the more complex and computationally expensive parallel multi-agent baseline. This suggests that simply presenting users with multiple, independent expert opinions may not necessarily lead to a better experience. For instance, as one evaluator mentioned, *"[The single agent's] strength [over the parallel multi-agent system] lies in its directness, providing immediate personalized information of BMI, clearly linking the user's goals, offering a simple and actionable first step for exercise, and prioritizing professional medical advice."* Without a clear conversational flow or organic synthesis between the different agent perspectives, in some cases users can find the response from the parallel multi-agent system confusing or disjointed.

Computational Cost. We also compare the computational cost across the three agent systems by the number of LLM calls and response time. Unsurprisingly, due to its simplicity, the single-agent PHIA has the lowest computational cost (average 3.0 calls per round, average response time 36.4 ± 1.6s), followed by PHA (average 6.5 calls per round, average response time 205.3 ± 15.0s), and parallel multi-agent system (average 7.0 calls per round, average response time 244.4 ± 11.4s). Due to its dynamic assignment, even with additional reflection stages, PHA requires less LLM calls and responses faster than the parallel multi-agent baseline. This highlights that PHA's intelligent orchestration not only delivers a superior user experience but also achieves it with greater computational efficiency than a static, parallel multi-agent approach.

7.3.2. Task 2: Multi-agent Conversation Experience for Experts

Evaluation Goal. Beyond the end-user perspective, this evaluation seeks to assess the technical and clinical validity of the agent's conversations from the viewpoint of health experts. The goal is to evaluate the soundness of the data analysis, the accuracy of the health information provided, the appropriateness of the coaching strategies, the coordination across three agents, as well as the overall quality of the conversation.

Evaluation Setup. We used the same 150 conversation scripts generated for the end-user evaluation. Five experts (4 females, 1 male, aged 25 - 64, 34.5 ± 5.7) with professional backgrounds in health coaching and digital health (three specializing in fitness, one in nutrition, and one in sports) reviewed all scripts while being blind to agent conditions. **Supplemental Section F.3.2** shows the detailed instruction for the expert evaluation. While the evaluation metrics covered similar domains as the end-user Task 1 (DS, DE, HC), the questions were framed from an expert standpoint, such as the clinical validity of the domain expert's advice, the correctness of the data science agent's analysis, and whether the health coach agent applied evidence-based motivational interviewing appropriately and safely (see questions in **Supplemental Table S30**). This expert-driven evaluation provides another layer of validation for the quality and reliability of the agent's responses. In total, the evaluation involves around 270 hours of human effort, resulting in 750 individual agent ratings (150 scripts × 5 experts) and 250 comparative rankings (50 personas × 5 experts).

Evaluation Results. The evaluation from health experts, summarized in **Figure 6(c)**, reveals a much more pronounced advantage for the PHA system than the results derived from end-users in **Section 7.3.1**. Coach experts, with their professional background, identified significant quality differences that were less apparent to end-users.

In the assessment of response quality, the gap between systems was salient. PHA received a positive overall quality rating from experts: 84.8±0.9%, a dramatic improvement over the parallel multi-agent (70.8±0.9%), and the single-agent system (65.2±0.8%, $p < 0.001$, GLMM with Gamma link function). Meanwhile, significant advantages of PHA were observed on all metrics across data science, domain knowledge, and coaching skill (see statistical details in **Supplemental Table S34**). These findings suggest that the two baseline systems, which lack PHA's dynamic agent task assignment and iterative reflection step, are more likely to produce responses that are factually shallow, clinically less precise, or contain inaccuracies that are readily apparent to a professional standard.

Such expert preference was confirmed in the final rankings. Experts ranked PHA with the best quality and as the most effective system for helping users achieve their goals in 80.0±2.5% of cases, more than five times of the preference for the parallel system (15.6±2.3% and 15.2±2.3% for Response-Quality and Effectiveness) and that of the single-agent (4.4±1.3% and 4.8±1.4%, $ps<0.001$, Friedman tests followed by post-hoc Wilcoxon Rank-sum tests with Holm-Bonferroni correction, **Supplemental Table S35**). Interestingly, a key discrepancy emerged between rater groups. While end-users rated the single-agent (PHIA) and parallel multi-agent baselines similarly, experts strongly disfavored the single agent, ranking at the bottom nearly three times as often (66.8±3.0% vs. 25.2±2.7% in Response-Quality). This suggests a difference in priorities: users may value a single, coherent narrative, while experts are better able to deconstruct the response and appreciate the specialized depth from each agent in the parallel system, even if the synthesis is imperfect. This discrepancy demonstrates that an expert evaluation can reveal shortcomings in technical depth that a user-focused assessment might overlook, again highlighting the importance of incorporating both expert and end-user perspectives in evaluation.

Experts' qualitative comments also complement these quantitative results. For instance, one expert noted *"[PHA] was a solid transcript where it was client-driven, collaborative and the outcome was a happy client with direction and a plan. [The single agent] was the opposite - no personalization and*

purely transactional... [The parallel multi agent baseline] had solid evidenced-based suggestions but failed to ask the client first which wasted time and removed the collaboration piece that is so critical with coaching." This demonstrates that the synergy between the DS, DE, and HC Agents within the PHA framework is not just a theoretical benefit but translates into a tangibly more effective, accurate, and professionally sound user experience.

Overarching Summary. After evaluating each single DS, DE, and HC agent independently, the final evaluation of our integrated PHA concluded the last piece of our evaluation framework and revealed that the whole is greater than the sum of its parts. Our orchestrated, collaborative framework significantly outperformed baselines. Ultimately, this comprehensive, multi-stage evaluation framework provides more than a performant system; it offers a validated blueprint for designing the next generation of personal health AI. Our findings champion a move away from monolithic models toward modular, multi-agent systems that emulate the collaborative and specialized nature of human expert teams, paving the way for health AI that is not only intelligent but also trustworthy, coherent, and genuinely helpful.

8. Discussion

Which Analyses Can the Data Science Agent Conduct?

The DS Agent was evaluated on its ability to translate ambiguous, health-related user queries into executable statistical analysis plans and corresponding code. The agent demonstrated significant proficiency in generating structured analysis plans, achieving a mean score of 75.6% on the plan auto-rater, compared to a baseline of 53.7%. This score reflects near expert-level performance in the planning stage. Specifically, the agent excelled at operationalizing ambiguous concepts, selecting appropriate analysis timeframes based on data availability and query context, identifying potential data sufficiency issues, and selecting appropriate statistical tests.

While the agent effectively generated structured analysis plans, performance was hindered by two key limitations. First, the plans often overlooked the crucial aspect of data distributions, a challenge stemming from the advanced statistical knowledge required, the difficulty of inferring distributions from limited personal data, and the complexity of integrating distributional considerations throughout the analysis. Second, translating plans into executable Python code, while achieving a 79.0% at the end, at times proved unreliable and inefficient. The agent frequently required multiple iterations of code generation, execution, and self-correction to produce working code, resulting in less than optimal latency (20-50 seconds). This iterative process, often necessary to resolve errors arising from complex multi-table operations like joins (e.g., incorrect keys, mismatched data types), highlights the need for substantial improvements in both the robustness of statistical reasoning and the reliability and efficiency of code generation.

Which Questions Can the Domain Expert Agent Accurately Answer?

The DE agent demonstrated improvements in delivering factual and trustworthy responses within the personal health and wellness domain, primarily due to its integration of factuality and citation tools. This allowed the agent to access and cite real-time information from authoritative sources, leading to quantifiable and statistically significant gains in both comprehensiveness and trustworthiness of responses as evaluated by clinicians (difference of 86.6% and 64.8% vs. baseline, respectively). Furthermore, the DE agent also demonstrated a strong ability to personalize responses, with the DE Agent's contextualization capabilities resulting in a 43.8% increase in user-rated Contextualization-and-Personalization win rate compared to the base model.

Despite these promising results, our analysis revealed several key limitations that warrant further investigation. First, the DE agent, leveraging a multi-step reasoning architecture, exhibited incomplete

reasoning chains in 8.5% of interactions, resulting in failure to provide a definitive conclusion or lacking sufficient supporting evidence. More importantly, its reliance on web search, while broadly useful, introduced vulnerabilities to conflicting information and highlighted a critical challenge in selecting authoritative sources. This issue of sourcing presents an opportunity not only to improve accuracy but also to educate users and enhance their health literacy. For instance, a future agent could explicitly contrast findings by stating, "While a general internet search may suggest X, the most reliable and up-to-date research indicates Y." This approach would transparently model the critical evaluation of health information for the user. Enabling such a feature requires refined tool-selection mechanisms, such as restricting the search tool to curated health domains or prioritizing queries to indexed scientific literature over general web searches.

In addition, the contextualizing and personalizing capability of the DE agent also introduced the risk of bias, as the model might inadvertently rely on stereotypes or heuristics when interpreting a user's health data. For instance, an agent could provide recommendations based on medical research that has historically underrepresented certain demographic groups, or make lifestyle suggestions that are not equitable across different socioeconomic backgrounds. Mitigating these risks is a critical area for future work. This includes performing rigorous algorithmic audits with diverse user personas to proactively uncover and address biased outputs, as well as refining the agent's information-retrieval process to prioritize more inclusive and representative data sources.

What Are the Important Characteristics of an Health Coach Agent?

Our evaluation reveals that the HC Agent, through its modular architecture, successfully navigates complex multi-turn dialogues. It demonstrates a superior ability to structure conversations, apply motivational interviewing techniques, and adapt to user feedback, leading to higher ratings from both end-users and experts. Beyond these capabilities, our findings also underscore a critical principle in conversational agent design: users prioritize core competency over conversational style. While stylistic elements such as a motivational tone, active listening cues, or client empowerment strategies can enhance the user experience, their positive impact is contingent upon the agent mastering the fundamentals of coaching. This foundation includes a deep understanding of the user's stated and unstated goals, a thorough exploration of their motivations, and skillful elicitation of relevant context (constraints, preferences, and prior attempts at solutions). Without this foundational understanding, even the most personable or empathetic agent will likely fail to provide valuable assistance. In fact, our study demonstrates that stylistic flourishes are often perceived negatively when the agent lacks core competencies. Users become frustrated when an agent appears to be "going through the motions" of a coaching conversation without actually providing personalized, actionable, and context-sensitive guidance. This highlights that stylistic elements must build upon a solid foundation of core coaching skills. The absence of competency causes users to view the agent as inauthentic, ineffective, and potentially a waste of their time.

Despite its strengths, our evaluation also identified a key limitation. Across both end-user and expert assessments, the HC Agent was rated lower than the baseline on the dimension of progress measurement. This suggests that the agent's current design, while excelling at initial goal setting and context gathering, is less optimized for the ongoing task of tracking and discussing progress towards those goals. Our iteration in PHA improves this capability by updating the prompt instruction, yet it can be further enhanced by additional modular designs in future iterations.

What Are the Benefits of the Personal Health Agent? How to Further Improve?

Our dynamic multi-agent framework (PHA) offers several key advantages. First, it enables a modular design, with each agent possessing clearly defined responsibilities and boundaries. This modularity facilitates both independent evaluation of individual agent capabilities and concurrent development of

specialized agentic modules (Sections 4, 5, 6). Second, our PHA addresses the inherent complexity of personal health inquiries, which often require integrating diverse forms of knowledge and reasoning, from interpreting extreme physiological values to extracting actionable insights and generating personalized recommendations. Our results demonstrate PHA's capacity as a conceptual model for this research to effectively synthesize these multidisciplinary skills, achieving a significant improvement in performance on complex health queries that require the integration of clinical knowledge and personal health data. Such advantages were validated through the perspectives of both end-users (advantage over baselines 1.6%-15.4%, average 4.6%) and experts (advantage over baselines 5.7%-39.0%, average 17.1%). Our multi-agent approach provides a robust and scalable foundation for building sophisticated and capable personal health agents.

Other than the directions of improvement on each single agent discussed above, future work can further explore more collaboration paradigms. While purely designed for research purposes, our PHA provides one potential pathway towards the vision of general personal health agents. It adopts a cooperative approach (Tran et al., 2025), where subagents collaborate together to address users' queries. Such an architecture is simpler to design and execute with clear goals. However, misaligned goals can cause inefficiencies among agents and one agent's failure can be amplified through cooperation. Another direction is competitive multi-agent systems, where sub-agents are pushed to compete against each other and promote adaptive strategies. For example, a "planner" agent might propose a new dietary adjustment, while an adversarial "critic" agent is incentivized to find potential flaws, such as nutrient deficiencies or conflicts with the user's stated preferences. To be beneficial, such a system requires robust adjudication mechanisms to resolve conflicts and synthesize a final recommendation that is safe and optimal.

Overall, the most promising frontier for research exploration may not be an exclusive choice between cooperation and competition, but in creating a dynamic pool of agents that can adapt their interaction patterns. It is important to note that clinical safety and regulatory analysis would need to be conducted before any real-world application. We envision a system where agents form temporary coalitions to solve novel problems, debate the merits of different health strategies with transparent reasoning, and evolve their roles as the user's own health journey unfolds. This paradigm shifts the goal from creating a mere question-answering tool to engineering a lifelong, personalized health agent as a co-pilot for users' daily experience.

Beyond agent architectural design, another crucial aspect for improvement lies in the agent's ability to adapt its interaction style to the user's dynamic psychological state, including personality traits, motivational drivers, and emotional context. This requires moving beyond static, one-size-fits-all interactions to an adaptive model that can dynamically select its conversational stance and motivational framing. For instance, one user might be more receptive to strategies that leverage accountability and principles of loss aversion, while another may respond better to supportive dialogue grounded in empathy and positive reinforcement. Furthermore, an individual's receptiveness to these strategies is not fixed; it can fluctuate based on their progress or other contextual factors. Successfully implementing such adaptability presents a significant research challenge, requiring sophisticated user modeling to accurately infer these user states and map them to optimal interaction paradigms.

Which User Journeys are Best Served?

Our analysis shows that the PHA demonstrates strong competency across all four CUJs, with particular strength over existing systems in personal health data interpretation (CUJ2) and generating wellness advice (CUJ3). The framework excels at identifying patterns in wearable data, contextualizing these findings against established health guidelines, and producing actionable, personalized recommendations.

However, addressing symptom assessment (CUJ4) presents a more complex and critical challenge. While PHA outperforms general-purpose models in providing basic symptom information (**Section 5.3.2**), its current architecture is not designed to offer definitive medical diagnoses or replace the judgment of a qualified healthcare professional. Yet, simply deferring this function to human clinicians overlooks a significant motivation for end-users seeking AI tools: the difficulty and delay in obtaining a diagnosis within the traditional healthcare system. For many end-users, AI represents a potential tool for making sense of their experience while they seek professional care.

This highlights a crucial opportunity for future development. The role of AI in CUJ4 is not to supplant the clinician, but to empower the patient throughout their diagnostic odyssey. In addition to further improving its diagnostic accuracy, future iterations could focus on helping users systematically log symptoms, identify patterns to discuss with a doctor, formulate relevant questions for appointments, and navigate the healthcare system more effectively. Pursuing this path requires careful design and a robust discussion of the ethical and policy implications to avoid the pitfalls of premature or inaccurate diagnostic suggestions. Ultimately, advancing CUJ4 adds an additional role of a personal health agent as a sophisticated preparatory and collaborative tool that helps bridge the gap between a patient's lived experience and a clinician's diagnostic expertise.

Risk and Ethical Considerations for Future Safeguards

The development and deployment of PHA introduce critical ethical considerations that demand robust safeguards. First, there is a significant risk of algorithmic bias. The underlying models may perpetuate or amplify existing health disparities if their training data underrepresents certain demographic groups, leading to less effective or inequitable recommendations. Future work must incorporate rigorous auditing for bias and develop fairness-aware mitigation strategies. Second, the ability of the PHA agent to generate and execute code on sensitive health data creates substantial security and privacy risks. A flawed script could lead to incorrect health insights or data breaches. This necessitates careful sandboxing of the execution environment, with strict limitations on capabilities (e.g., preventing network access) and validation of all generated code. Third, a key challenge is managing user reliance and ensuring safety. Users may over-rely on the agent for issues that demand professional medical consultation. The system must therefore include clear, persistent disclaimers and mechanisms to detect high-risk queries, proactively directing users to clinicians. Finally, the deployment of such systems raises significant regulatory considerations. Ensuring compliance with data privacy regulations (e.g., HIPAA, GDPR), emerging AI regulations, establishing clear guidelines for clinical validation, and addressing liability are essential prerequisites for responsible implementation and require a thorough regulatory assessment.

Limitations

Despite the promising capabilities of the PHA framework, several limitations warrant consideration and future research. First, while our evaluation demonstrates short-term effectiveness in providing personalized health information and recommendations, we have not yet conducted a longitudinal study to assess the long-term impact on user behavior change through multi-session interactions. Sustained engagement and adherence to personalized health plans are critical for achieving meaningful improvements in health outcomes, and future work will need to rigorously evaluate PHA's ability to promote and maintain positive behavioral modifications over extended periods. Second, our evaluation methodology relies heavily on autoraters. While this approach allows for faster and scalable evaluation, care must be taken to account for autorater biases such as sensitivities to dataset ordering, formatting, length, and source. Third, the computational cost of the current PHA implementation, with its three specialized agents and an orchestration module, is substantial. Although PHA shows advantages over the parallel multi-agent baseline, compared to the single-agent PHIA baseline, PHA involves more number of LLM calls (3.0 vs 6.5 calls per query on average) and increases the latency

(36.4s vs. 244.4s response time on average, see **Section 7.3.1**). This presents a barrier to scalability and widespread adoption, particularly in resource-constrained settings. Future research will focus on optimizing the agent architecture, exploring more efficient inference methods, and potentially leveraging specialized hardware to reduce both latency and cost, making the system more accessible and practical for real-world deployments.

9. Conclusion

Creating AI systems that can not only analyze and interpret complex personal health data but also provide empathetic and actionable wellness advice has been a longstanding grand challenge in AI. Our approach deconstructs this problem into three core capabilities with three sub-agents: Data Science, Domain Expert, and Health Coach, that excel in their respective roles. Our integrated multi-agent framework PHA orchestrates these distinct agents to deliver contextualized, personalized, and evidence-based responses. Through extensive evaluation across 10 benchmark tasks and over 1100 human evaluation involving both end-users and health experts, we demonstrate the advantage of each individual agents and the multi-agent framework over baseline solutions. This work lays the foundation towards our vision of a personal health agent that can support human health expertise and is accessible to everyone in the future.

Acknowledgements

We would like to thank Andrew Mai, Florence Gao, Vaishnavi Vinod Kumar, and Vibhati Sharma for coordinating human evaluations, as well as all expert and end-user/expert raters who evaluated model outputs.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to {aliheydari, dmcduff, xliucs, orson}@google.com.

References

T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Trans. Assoc. Comput. Linguist.*, 4:463–476, Aug. 2016.

E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit. Med.*, 8(1):274, May 2025.

G. Bai, J. Liu, X. Bu, Y. He, J. Liu, Z. Zhou, Z. Lin, W. Su, T. Ge, B. Zheng, and W. Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. Feb. 2024.

E. Basar, I. Hendrickx, E. Krahmer, G.-J. Bruijn, and T. Bosse. To what extent are large language models capable of generating substantial reflections for motivational interviewing counseling chatbots? a human evaluation. In N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, and H. A. Schwartz, editors, *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 41–52, Stroudsburg, PA, USA, 2024. ACL.

M. Benary, X. D. Wang, M. Schmidt, D. Soll, G. Hilfenhaus, M. Nassir, C. Sigler, M. Knödler, U. Keller, D. Beule, U. Keilholz, U. Leser, and D. T. Rieke. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open*, 6(11):e2343689, Nov. 2023.

E. Bennett, R. Alpert, and A. Goldstein. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308, Sept. 1954.

BoardVitals. Cardiology board review. https://www.boardvitals.com/cardiology-board-review/, 2024. Accessed: 15 July 2025.

J. Chen, S. Saha, and M. Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. *arXiv [cs.LG]*, July 2021.

L. Cheng, X. Li, and L. Bing. Is GPT-4 a good data analyst? In *The 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023.

C.-H. Chiang and H.-Y. Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

Data Commons. Data commons.
urlhttps://datacommons.org, 2025. Accessed on July 7, 2025.

Y. Deng, L. Liao, Z. Zheng, G. H. Yang, and T.-S. Chua. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 2022, pages 807–818, New York, NY, USA, July 2024. ACM.

V. Dhar. Data science and prediction. *Commun. ACM*, 56(12):64–73, Dec. 2013.

V. Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv [cs.AI]*, Mar. 2023.

G. T. Doran. There's a smart way to write managements's goals and objectives. *Management review*, 70(11), 1981.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, June 2024.

Fitbit Community. Fitbit labs: Testing new, experimental health and fitness capabilities in the fitbit app, 2024. URL https://community.fitbit.com/t5/The-Pulse-Fitbit-Community-Blog/Fitbit-Labs-Testing-new-experimental-health-amp-fitness-capabilities-in-the/ba-p/5675311.

J. A. Fodor. *The modularity of mind*. A Bradford Book. Bradford Books, Cambridge, MA, Jan. 1983.

H. Fraser, D. Crossland, I. Bacher, M. Ranney, T. Madsen, and R. Hilliard. Comparison of diagnostic and triage accuracy of ada health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR MHealth UHealth*, 11(1): e49995, Oct. 2023.

Google. Google Web Search. https://www.google.com, 2025.

J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk, A. Vahdat, P. Kohli, Y. Matias, A. Carroll, K. Kulkarni, N. Tomasev, Y. Guan, V. Dhillon, E. D. Vaishnav, B. Lee, T. R. D. Costa, J. R. Penadés, G. Peltz, Y. Xu, A. Pawlosky, A. Karthikesalingam, and V. Natarajan. Towards an AI co-scientist. Feb. 2025.

G. Graham, N. Goren, V. Sounderajah, and K. DeSalvo. Information is a determinant of health. *Nat. Med.*, 30(4):927–928, Apr. 2024.

K. Gu, M. Grunde-McLaughlin, A. McNutt, J. Heer, and T. Althoff. How do data analysts respond to AI assistance? a wizard-of-oz study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 38, pages 1–22, New York, NY, USA, May 2024a. ACM.

K. Gu, R. Shang, R. Jiang, K. Kuang, R.-J. Lin, D. Lyu, Y. Mao, Y. Pan, T. Wu, J. Yu, Y. Zhang, T. M. Zhang, L. Zhu, M. A. Merrill, J. Heer, and T. Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971, Stroudsburg, PA, USA, Nov. 2024b. Association for Computational Linguistics.

K. Gu, Z. Zhang, K. Lin, Y. Zhang, A. Paruchuri, H. Yu, M. Kazemi, K. Ayush, A. A. Heydari, M. A. Xu, G. Narayanswamy, Y. Liu, M.-Z. Poh, Y. Yang, M. Malhotra, S. Patel, H. Palangi, X. Xu, D. McDuff, T. Althoff, and X. Liu. Radar: Benchmarking language models on imperfect tabular data, 2025a. URL https://arxiv.org/abs/2506.08249.

K. Gu, Z. Zhang, K. Lin, Y. Zhang, A. Paruchuri, H. Yu, M. Kazemi, K. Ayush, A. A. Heydari, M. A.

Xu, et al. Radar: Benchmarking language models on imperfect tabular data. *arXiv preprint arXiv:2506.08249*, 2025b.

S. Guo, C. Deng, Y. Wen, H. Chen, Y. Chang, and J. Wang. Ds-agent: automated data science by empowering large language models with case-based reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

J. Hettema, J. Steele, and W. R. Miller. Motivational interviewing. *Annu. Rev. Clin. Psychol.*, 1(1): 91–111, Apr. 2005.

S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, New York, NY, USA, Sept. 2019. ACM.

S. Hong, Y. Lin, B. Liu, B. Liu, B. Wu, C. Zhang, C. Wei, D. Li, J. Chen, J. Zhang, J. Wang, L. Zhang, L. Zhang, M. Yang, M. Zhuge, T. Guo, T. Zhou, W. Tao, X. Tang, X. Lu, X. Zheng, X. Liang, Y. Fei, Y. Cheng, Z. Gou, Z. Xu, and C. Wu. Data interpreter: An LLM agent for data science. Feb. 2024.

X. Hu, Z. Zhao, S. Wei, Z. Chai, Q. Ma, G. Wang, X. Wang, J. Su, J. Xu, M. Zhu, Y. Cheng, J. Yuan, J. Li, K. Kuang, Y. Yang, H. Yang, and F. Wu. InfiAgent-DABench: Evaluating agents on data analysis tasks. In *International Conference on Machine Learning*, pages 19544–19572. PMLR, July 2024a.

X. Hu, Z. Zhao, S. Wei, Z. Chai, Q. Ma, G. Wang, X. Wang, J. Su, J. Xu, M. Zhu, Y. Cheng, J. Yuan, J. Li, K. Kuang, Y. Yang, H. Yang, and F. Wu. Infiagent-dabench: evaluating agents on data analysis tasks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Y. Ikemoto, V. Asawavetvutt, K. Kuwabara, and H.-H. Huang. Conversation strategy of a chatbot for interactive recommendations. In *Intelligent Information and Database Systems*, Lecture notes in computer science, pages 117–126. Springer International Publishing, Cham, 2018.

J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. A survey on large language models for code generation. *arXiv [cs.CL]*, June 2024.

M. Jörke, S. Sapkota, L. Warkenthien, N. Vainio, P. Schmiedmayer, E. Brunskill, and J. Landay. Supporting physical activity behavior change with LLM-based conversational agents. May 2024.

E. Jun, A. Seo, J. Heer, and R. Just. Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships. *arXiv [cs.AI]*, Jan. 2022.

M. Jörke, S. Sapkota, L. Warkenthien, N. Vainio, P. Schmiedmayer, E. Brunskill, and J. A. Landay. GPTCoach: Towards LLM-based physical activity coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–46, New York, NY, USA, Apr. 2025. ACM.

Z. Kanjee, B. Crowe, and A. Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80, July 2023.

J. Khasentino, A. Belyaeva, X. Liu, Z. Yang, N. A. Furlotte, C. Lee, E. Schenck, Y. Patel, J. Cui, L. D. Schneider, R. Bryant, R. G. Gomes, A. Jiang, R. Lee, Y. Liu, J. Perez, J. K. Rogers, C. Speed, S. Tailor, M. Walker, J. Yu, T. Althoff, C. Heneghan, J. Hernandez, M. Malhotra, L. Stern, Y. Matias, G. S. Corrado, S. Patel, S. Shetty, J. Zhan, S. Prabhakara, D. McDuff, and C. Y. McLean. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, Aug 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03888-0. URL https://www.nature.com/articles/s41591-025-03888-0.

J.-W. Kim, B. Ryu, S. Cho, E. Heo, Y. Kim, J. Lee, S. Y. Jung, and S. Yoo. Impact of personal health

records and wearables on health outcomes and patient response: Three-arm randomized controlled trial. *JMIR MHealth UHealth*, 7(1):e12070, Jan. 2019.

Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, C. Breazeal, and H. W. Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *Neural Inf Process Syst*, 37:79410–79452, Apr. 2024.

Y. Kim, Z. Hu, H. Jeong, E. Park, S. S. Li, C. Park, S. Xiong, M. Lu, H. Lee, X. Liu, D. McDuff, C. Breazeal, S. Tulebaev, and H. W. Park. BehaviorSFT: Behavioral token conditioning for clinical agents across the proactivity spectrum. *arXiv [cs.CL]*, May 2025a.

Y. Kim, H. Jeong, C. Park, E. Park, H. Zhang, X. Liu, H. Lee, D. McDuff, M. Ghassemi, C. Breazeal, S. Tulebaev, and H. W. Park. Tiered agentic oversight: A hierarchical multi-agent system for AI safety in healthcare. *arXiv [cs.AI]*, June 2025b.

Y. Kim, T. Kim, W. Kang, E. Park, J. Yoon, D. Lee, X. Liu, D. McDuff, H. Lee, C. Breazeal, and H. W. Park. VocalAgent: Large language models for vocal health diagnostics with safety-aware evaluation. *arXiv [cs.SD]*, May 2025c.

S. H. Lee. Natural language generation for electronic health records. *NPJ Digit. Med.*, 1(1):63, Nov. 2018.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *arXiv [cs.AI]*, Mar. 2023.

J. Li, N. Huo, Y. Gao, J. Shi, Y. Zhao, G. Qu, B. Qin, X. Li, C. Ma, J.-G. Lou, and R. Cheng. Benchmarking intelligent LLM agents for conversational data analysis. Oct. 2024.

J. Liu, N. Boukhelifa, and J. R. Eagan. Understanding the role of alternatives in data analysis practices. *IEEE Trans. Vis. Comput. Graph.*, 26(1):66–76, Jan. 2020a.

J. Liu, N. Boukhelifa, and J. R. Eagan. Understanding the role of alternatives in data analysis practices. *IEEE Trans. Vis. Comput. Graph.*, 26(1):66–76, Jan. 2020b.

Y. Liu, T. Althoff, and J. Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. *arXiv [cs.HC]*, Oct. 2019.

Y. Liu, T. Althoff, and J. Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020c. ACM.

Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.*, 26(6):900–908, June 2020d.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-refine: Iterative refinement with self-feedback. *arXiv [cs.CL]*, Mar. 2023.

M. B. Madson, R. S. Mohn, A. Zuckoff, J. A. Schumacher, J. Kogan, S. Hutchison, E. Magee, and B. Stein. Measuring client perceptions of motivational interviewing: factor analysis of the client evaluation of motivational interviewing scale. *J. Subst. Abuse Treat.*, 44(3):330–335, Mar. 2013.

M. Maguire and B. Delahunt. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J*, 9(3), Oct. 2017.

N. Mallinar, A. A. Heydari, X. Liu, A. Z. Faranesh, B. Winslow, N. Hammerquist, B. Graef, C. Speed, M. Malhotra, S. Patel, J. L. Prieto, D. McDuff, and A. A. Metwally. A scalable framework for evaluating health language models, 2025.

S. Mantena, A. Johnson, M. Oppezzo, N. Schuetz, A. Tolas, R. Doijad, C. M. Mattson, A. Lawrie, M. Ramirez-Posada, E. Linos, A. C. King, F. Rodriguez, D. S. Kim, and E. A. Ashley. Fine-tuning large language models in behavioral psychology for scalable physical activity coaching. Feb. 2025.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, L. Hou, Y. Cheng, Y. Liu, S. Sara Mahdavi, S. Prakash, A. Pathak, C. Semturs, S. Patel, D. R. Webster, E. Dominowska, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, J. Sunshine, A. Karthikesalingam, and V. Natarajan. Towards accurate differential diagnosis with large language models. Nov. 2023.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, and others. Towards accurate differential diagnosis with large language models. *Nature*, 2025.

M. Mehta, D. Caperton, K. Axford, L. Weitzman, D. Atkins, V. Srikumar, and Z. Imel. Psychotherapy is not one thing: Simultaneous modeling of different therapeutic approaches. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 47–58, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

J. Mercado, I. E. Espinosa-Curiel, and J. Martínez-Miranda. Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *J. Med. Internet Res.*, 25(1):e52097, Dec. 2023.

M. A. Merrill, A. Paruchuri, N. Rezaei, G. Kovacs, J. Perez, Y. Liu, E. Schenck, N. Hammerquist, J. Sunshine, S. Tailor, K. Ayush, H.-W. Su, Q. He, C. Y. McLean, M. Malhotra, S. Patel, J. Zhan, T. Althoff, D. McDuff, and X. Liu. Transforming wearable data into health insights using large language model agents. June 2024a.

M. A. Merrill, M. Tan, V. Gupta, T. Hartvigsen, and T. Althoff. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3512–3533, Stroudsburg, PA, USA, Nov. 2024b. Association for Computational Linguistics.

A. A. Metwally, A. A. Heydari, D. McDuff, A. Solot, Z. Esmaeilpour, A. Z. Faranesh, M. Zhou, D. B. Savage, C. Heneghan, S. Patel, C. Speed, and J. L. Prieto. Insulin resistance prediction from wearables and routine blood biomarkers. *arXiv [cs.LG]*, Apr. 2025.

T. Munder, F. Wilmers, R. Leonhart, H. W. Linster, and J. Barth. Working alliance Inventory-Short revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clin. Psychol. Psychother.*, 17(3):231–239, May 2010.

NCBI. Ncbi. https://www.ncbi.nlm.nih.gov/, 2025.

H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv [cs.CL]*, Nov. 2023.

H. Nori, N. Usuyama, N. King, S. M. McKinney, X. Fernandes, S. Zhang, and E. Horvitz. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv [cs.CL]*, Nov. 2024.

Oura Team. Introducing oura advisor: Your ai-powered personal health companion. https://ouraring.com/blog/oura-advisor/, 2025. Accessed: 2025-07-10.

A. Palepu, V. Liévin, W.-H. Weng, K. Saab, D. Stutz, Y. Cheng, K. Kulkarni, S. S. Mahdavi, J. Barral, D. R. Webster, K. Chou, A. Hassidim, Y. Matias, J. Manyika, R. Tanno, V. Natarajan, A. Rodman, T. Tu, A. Karthikesalingam, and M. Schaekermann. Towards conversational AI for disease management. *arXiv [cs.CL]*, Mar. 2025.

T. T. Quoc, D. H. Minh, T. Q. Thanh, and A. Nguyen-Duc. An empirical study on self-correcting large language models for data science code generation. *arXiv [cs.SE]*, Aug. 2024.

M. Ringeval, G. Wagner, J. Denford, G. Paré, and S. Kitsiou. Fitbit-based interventions for healthy lifestyle outcomes: Systematic review and meta-analysis. *J. Med. Internet Res.*, 22(10):e23954, Oct. 2020.

K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, J. Z. Chaves, S.-Y. Hu, M. Schaekermann, A. Kamath, Y. Cheng, D. G. T. Barrett, C. Cheung, B. Mustafa, A. Palepu, D. McDuff, L. Hou, T. Golany, L. Liu, J.-B. Alayrac, N. Houlsby, N. Tomasev, J. Freyberg, C. Lau, J. Kemp, J. Lai, S. Azizi, K. Kanada, S. Man, K. Kulkarni, R. Sun, S. Shakeri, L. He, B. Caine, A. Webson, N. Latysheva, M. Johnson, P. Mansfield, J. Lu, E. Rivlin, J. Anderson, B. Green, R. Wong, J. Krause, J. Shlens, E. Dominowska, S. M. A. Eslami, C. Cui, O. Vinyals, K. Kavukcuoglu, J. Manyika, J. Dean, D. Hassabis, Y. Matias, D. Webster, J. Barral, G. Corrado, C. Semturs, S. S. Mahdavi, J. Gottweis, A. Karthikesalingam, and V. Natarajan. Capabilities of gemini models in medicine. Apr. 2024.

J. Schaffer. What not to multiply without necessity. *Australas. J. Philos.*, 93(4):644–664, Oct. 2015.

A. Sharma, K. Rushton, I. Lin, D. Wadden, K. Lucas, A. Miner, T. Nguyen, and T. Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

A. Sharma, K. Rushton, I. W. Lin, T. Nguyen, and T. Althoff. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 21, pages 1–29, New York, NY, USA, May 2024. ACM.

U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208 – 1214, 2020. URL https://api.semanticscholar.org/CorpusID:220810602.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, Aug. 2023.

V. Srinivas, X. Xu, X. Liu, K. Ayush, I. Galatzer-Levy, S. Patel, D. McDuff, and T. Althoff. Substance over style: Evaluating proactive conversational coaching agents. Mar. 2025.

StatPearls Publishing. American board of internal medicine: Endocrinology, diabetes, & metabolism exam. https://www.statpearls.com/boardreview/Endocrinology, 2024. Accessed: 15 July 2025.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*, 11(5):702–712, Sept. 2016a.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016b.

X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 599–621, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

The Gemini Team. Gemini: A family of highly capable multimodal models. Dec. 2023.

K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv [cs.AI]*, Jan. 2025.

T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, S. Azizi, K. Singhal, Y. Cheng, L. Hou, A. Webson, K. Kulkarni, S. Sara Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, and V. Natarajan. Towards conversational diagnostic AI. Jan. 2024.

J.-P. Vert. How will generative AI disrupt data science in drug discovery? *Nat. Biotechnol.*, 41(6): 750–751, June 2023.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, Sept. 2022.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, Sept. 2023.

WHOOP. Introducing whoop coach, powered by openai. https://www.whoop.com/us/en/thelocker/introducing-whoop-coach-powered-by-openai/, 2023. Accessed: 2025-07-10.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv [cs.AI]*, Aug. 2023.

X. Wu, R. Zheng, J. Sha, T.-L. Wu, H. Zhou, T. Mohan, K.-W. Chang, N. Peng, and H. Huang. DACO: Towards Application-Driven and comprehensive data analysis via code generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Nov. 2024a.

X. Wu, R. Zheng, J. Sha, T.-L. Wu, H. Zhou, M. Tang, K.-W. Chang, N. Peng, and H. Huang. DACO: Towards application-driven and comprehensive data analysis via code generation. *arXiv [cs.CL]*, Mar. 2024b.

X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.

L. Yang, S. Xu, A. Sellergren, T. Kohlberger, Y. Zhou, I. Ktena, A. Kiraly, F. Ahmed, F. Hormozdiari, T. Jaroensri, E. Wang, E. Wulczyn, F. Jamil, T. Guidroz, C. Lau, S. Qiao, Y. Liu, A. Goel, K. Park, A. Agharwal, N. George, Y. Wang, R. Tanno, D. G. T. Barrett, W.-H. Weng, S. S. Mahdavi, K. Saab,

T. Tu, S. R. Kalidindi, M. Etemadi, J. Cuadros, G. Sorensen, Y. Matias, K. Chou, G. Corrado, J. Barral, S. Shetty, D. Fleet, S. M. A. Eslami, D. Tse, S. Prabhakara, C. McLean, D. Steiner, R. Pilgrim, C. Kelly, S. Azizi, and D. Golden. Advancing multimodal medical capabilities of gemini. *arXiv [cs.CV]*, May 2024.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

P. Yin, W.-D. Li, K. Xiao, A. K. Rao, Y. Wen, K. Shi, J. Howland, P. Bailey, M. Catasta, H. Michalewski, O. Polozov, and C. Sutton. Natural language to code generation in interactive data science notebooks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, July 2023.

D. Zhang, S. Zhoubian, M. Cai, F. Li, L. Yang, W. Wang, T. Dong, Z. Hu, J. Tang, and Y. Yue. DataSciBench: An LLM agent benchmark for data science. Feb. 2025.

Y. Zhang, Q. Jiang, X. XingyuHan, N. Chen, Y. Yang, and K. Ren. Benchmarking data science agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5677–5700, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Neural Inf Process Syst*, abs/2306.05685:46595–46623, June 2023.

N. S. Zheng, J. Annis, H. Master, L. Han, K. Gleichauf, J. H. Ching, M. Nasser, P. Coleman, S. Desine, D. M. Ruderfer, J. Hernandez, L. D. Schneider, and E. L. Brittain. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the all of us research program. *Nat. Med.*, 30(9):2648–2656, Sept. 2024.

Supplementary Material

A. User-Centered Design & User Journey Details

A.1. User Journeys Breakdown with Agent Assignment

Table S1 | Personal Health Queries and Agent Assignment. Our user-centered design process revealed four major critical user journeys (CUJs) and 26 most common types of user queries, together with query examples. For each CUJ and user type, a group of researchers and health experts design workflows to involve appropriate members of the three agent team (DS: Data Science, DE: Domain Expert, HC: Health Coach) together with an agent collaboration workflow.

Category	User Purpose	Question Types (I want...)	Question Examples	Sub-Agents	Agent Workflow
Health knowledge (CUJ1)	I'm here to understand and discover health knowledge	...to understand x	What is HRV?	DE	DE agent answers based on internal and external knowledge
		...to find a fact about x	How long is strep contagious?	DE	
		...to compare x and y	What is the difference between an HRV & RHR?	DE	
		...the pros and cons of x	What are the pros and cons of HIIT?	DE	
		...the news about x	What's the latest research on sleep apnea?	DE	
Personal data (CUJ2)	I'm here to understand my wearable data	...to see my best/worst x data	What's my longest sleep duration last month?	DS	DS agent answers by computation w/ results.
		...to see or compare my x data for a time period y	What was my average heart rate during the Euro Cup?	DS	
		...to identify meaningful changes in x	Can you review my HRV data from this month and highlight any unusual occurrences?	DS, DE	DS agent does the computation and DE agent interprets the results and answer
		...to understand what x data says about my health	What is my current fitness level? How has my sleep been?	DS, DE	
		...to compare my x with other people/clinical guidelines	How does my exercise compare to pregnant women my age?	DS, DE	
		...to summarize x time-series data	Give me an overview of my physical activity records since I started my training.	DS, DE	

Table S1 – continued from previous page

Category	User Purpose	Question Types (I want...)	Question Examples	Sub-Agents	Agent Workflow
	I'm here to understand my health record	...to answer a question from x	Which lab values in my lipid panel should I pay attention to and why?	DE	DE agent interprets, reasons, and summarizes personal data. No complex computation is needed, so DS agent is not involved.
		...to summarize x (at certain timestamp)	Can you help me understand this blood report? Do I sleep better on days where I run for an hour?	DE	
	I'm here to understand relationships between my health data	...to see relationships between x and y in the same source (e.g., Fitbit)	Could you maybe explain why my steps fluctuate from weekdays to weekends?	DS	DS agent answers by computation w/ results.
		...to understand why x changed	Does my daily exercise in the past six months make my blood healthier?	DS, DE	DS agent does the computation and DE agent interprets the results and answer
		...to compare x and y across sources	My throat is itchy and I have trouble swallowing. What might be going on?	DS, DE	
Wellness advice (CUJ3)	I'm here to get wellness advice	...advice on x (e.g., improve / change) in general	Give me tips to reduce smartphone usage?	HC	HC agent provides wellness advice based on users' needs
		...advice on x (e.g., improve / change) based on personal data.	How do I improve my deep sleep?	DS, HC	DS agent does related analysis, and HC agent provides personalized wellness advice
		...to identify areas for improvement	What are my biggest problem areas in my exercise routine?	DS, DE, HC	DS agent analyzes personal data. DE agent interprets the results, and HC agent walks through and makes plans with users

Table S1 – continued from previous page

Category	User Purpose	Question Types (I want...)	Question Examples	Sub-Agents	Agent Workflow
	I'm here to make progress towards my goals	...to set an appropriate goal or task for x	How much exercise should I be doing to get rid of belly fat?	HC	HC agent sets goals and plans based on users preference and needs.
		...to understand how my behaviors move me toward my goal for x	What are some of the things I have done that have the biggest impact on my blood pressure/sleep?	DE, HC	DE agent provides domain-specific knowledge; HC agent makes plans based on users' preference and needs
		...plans to achieve my x goal	Give me a 8 week 10k training plan for pregnant women.	DS, DE, HC	DS agent analyzes personal data when appropriate. DE agent provides expert-level interpretation and advice, if needed. HC agent walks through with users to set goals and make plans.
		...to adjust my plans or goals based on changing needs	It's raining. Adjust my workout to be indoors today.	DS, DE, HC	
		...to track progress against my goals for x	Based on my training pace, am I on track for completing a marathon in under 3 hours?	DS, DE, HC	
Personal medical symptoms (CUJ4)	I'm here to assess my symptoms	...to explore x symptoms that I'm feeling	How can I tell if I have high blood pressure?	DE	DE agent diagnoses based on information provided
		...advice on a health symptom	How do I improve hypertension?	DS, DE	DS agent does related computation when appropriate, DE agent analyzes the results and provides advice

A.2. User-Centered Survey Details

Table S2 | End-User Survey for Ask Health Fitbit Labs. This survey was delivered to end-users who interacted with Ask Health Fitbit Labs (N=555) to ask about their personal health goals and sentiments when interacting with the agent.

Question Category	Question Body	Answer Options
Engagement / Retention	How often did you use Insights Explorer in the past week?	• Never • Once a week • 2-3 times a week • 4-6 times a week • Every day of the week
	During the past few weeks, did your usage of Insights Explorer increase, decrease, or remain the same?	• My usage increased • My usage decreased • My usage remained the same • I don't remember
	[Branch: If *decreased*] Please select any reasons that contributed to your decreased usage over time.	[multi-select, randomized order] • Forgot about it • Unsure what to ask about • I can easily get the information I'm looking for elsewhere • I can get better quality answers elsewhere • I didn't trust the information it provided • It was too limited in what it could answer (e.g., types of information) • Response not useful did not address my query, incomplete, etc • Content was made up or factually incorrect • Response was too slow • Other [open ended]
	[Branch: If *increased* or *stayed the same*] What was the main functionality or benefits that you continued using Insights Explorer for?	[open-text]

Table S2 – continued from previous page

Question Category	Question Body	Answer Options
	Overall, how satisfied or dissatisfied are you with Insights Explorer?	• Very dissatisfied • Somewhat dissatisfied • Neither satisfied nor dissatisfied • Somewhat satisfied • Very satisfied
	Indicate any aspects of Insights Explorer that have NOT met your expectations:	[multi-select, randomized order] • It wasn't able to answer my questions • Too limited in types of data I ask about (e.g., weight, exercise) • The information was obvious or already available in the app • Not able to ask about recent or real-time data • I didn't know how to use it, or what to use it for • It was not accurate or factually correct • The answers were poor quality or not useful • It didn't remember what I had previously asked • Other [open text]
Satisfaction	Which aspects of Insights Explorer, if any, delighted you?	[multi-select, randomized order] • charts • Custom • Correlations across my behaviors and health metrics • Trends in my health metrics • Ability to provide insights quickly without Browse the app • Comparing my health metrics to my demographic and across time • Helped me understand my health and wellness better • [Open-ended] • None of the above
	How satisfied are you with the suggested questions shown to you in Insights Explorer?	• Very dissatisfied • Somewhat dissatisfied • Neither satisfied nor dissatisfied • Somewhat satisfied • Very satisfied
	Please explain why you were satisfied or dissatisfied with the suggested questions.	[open ended]
	How often were you satisfied with the quality of answers to your questions?	• Never • Rarely • Sometimes • Often • Always
	Please explain why you were satisfied or dissatisfied with the quality of answers to your questions.	[open ended]
	On a scale from 0-10, how likely are you to recommend Insights Explorer to a friend or colleague?	0 (Not at all likely) - 10 (Extremely likely)

Table S2 – continued from previous page

Question Category	Question Body	Answer Options
Health Impact & Literacy	How much do you agree with this statement: "Using Insights Explorer helped me better understand my health and fitness."	• Strongly disagree • Disagree • Neither agree nor disagree • Agree • Strongly Agree
CUJ Identification	What are you hoping to achieve by using Insights Explorer? Please select all that apply.	[multi-select, randomized order] • Faster/easier way to get my Fitbit data • Connect the dots between my data points (e.g., how does my sleep impact my activity levels) • Understand how I fit in with the broader Fitbit population (e.g., do I walk more or less than other women my age) • Make comparisons of specific time periods (e.g., when I was ill vs. healthy) • Get recommendations based on my data (e.g., should I take it easy today?) • Get recommendations for personalized goals (e.g., what should my daily step count goal be?) • Learn more about Fitbit metrics and data types (e.g., how is the readiness score computed?) • Learn more about general health questions (e.g., what does it mean to have high cholesterol?) • Other [open ended]
	[For each goal selected] Please rate your satisfaction with your ability to do each of the following using Insights Explorer [show each goal before the rating scale]:	• Very dissatisfied • Somewhat dissatisfied • Neither satisfied nor dissatisfied • Somewhat satisfied • Very satisfied
Innovation	How technologically advanced does this feature seem?	• Extremely advanced • Very advanced • Somewhat advanced • Not very advanced • Not at all advanced
	How unique does this feature seem?	• Extremely unique • Very unique • Somewhat unique • Not very unique • Not at all unique
	How useful does this feature seem?	• Extremely useful • Very useful • Somewhat useful • Not very useful • Not at all useful

Table S2 – continued from previous page

Question Category	Question Body	Answer Options
	From the list below, which word best describes Insight Explorer?	• Annoying • Exciting • Intuitive • Boring • Fresh • Irrelevant • Busy • Friendly • Patronizing • Creative • Helpful • Professional • Confusing • Impressive • Trustworthy • Cutting-edge • Innovative • Unhelpful • Delightful • Intimidating • Unprofessional • Empowering
Trust	To what extent do you trust Insights Explorer?	• Not at all • Slightly • Moderately • Very • Completely
	How credible do you feel the information provided by Insights Explorer is?	• Not at all credible • Slightly credible • Moderately credible • Very credible • Extremely credible
Experience with other products	Apart from Insights Explorer, please select any of the following Generative AI products, features, or tools that you have personal experience using.	[multi-select, randomized order] • Gemini • ChatGPT • Coda OpenAI • DALL-E2 • Microsoft Co-Pilot • Mid Journey • Notion AI • Stable Diffusion • Another AI feature/capability/tool • None of these

B. Evaluation Dataset Preparation

B.1. Description of WEAR-ME Data Categories

B.1.1. Demographics

The Demographics category includes background information used to characterize the study population. This encompasses core attributes such as the participant's year of birth (BIRTH_YEAR_DATA), gender (GENDER_DATA), self-identified ethnicity (Ethnicity), and marital status (marital_status).

B.1.2. Anthropometrics

The anthropometrics category consists of physical body measurements. These data include objectively measured height (HEIGHT_DATA) and weight (WEIGHT_DATA), which are supplemented by self-reported values for weight (weight) and waist circumference (waist_circumference).

B.1.3. Socioeconomic Status

The Socioeconomic Status group provides insight into the participants' social and economic context, which can be a key determinant of health. This includes variables such as self-reported annual household income (household_income), current employment status (employment_status), and the highest level of education achieved (highest_education).

B.1.4. Female & Reproductive Health

This category contains data specific to the physiological and health status of female participants. It covers information regarding current menstruation status (menstruation), the self-reported regularity of menstrual periods (regular_periods), and the use of contraceptives (contraceptives).

B.1.5. Lifestyle & Behavior (Baseline)

This group captures a comprehensive snapshot of self-reported habits, behaviors, and perceptions *at the start* of the study. This includes data on physical activity (aerobic_before', strength_before'), dietary patterns (diet', eat_out'), alcohol and tobacco use (alcohol, smoker), typical sleep patterns (sleep_hours, sleep_quality), and subjective assessments of health (overall_health_before).

B.1.6. Lifestyle & Behavior (Follow-up)

This category includes variables that reassess key habits and perceptions after the study's primary observation period. This allows for the analysis of changes in behaviors such as aerobic activity (aerobic_after), eating habits (eating_after), and self-rated overall health (overall_health_after).

B.1.7. Digital Biomarkers (Wearables)

This group of features include a rich set of high-frequency physiological and behavioral metrics passively collected from wearable sensors. This includes detailed activity tracking, such as daily steps (DAILY_STEPS_DATA) and active zone minutes (DAILY_ACTIVE_ZONE_MINUTES_DATA), alongside comprehensive sleep analysis, covering sleep duration (DAILY_SLEEP_DURATION_DATA), time in sleep stages (DAILY_TIME_IN_SLEEP_STAGES_DATA), and a composite sleep score (SLEEP_SCORE_DATA). The category also contains key cardiovascular signals like daily resting heart rate (DAILY_RESTING_HEART_RATE_DATA) and heart rate variability (DAILY_HEART_RATE_VARIABILITY_DATA), as well as other physiological measures like respiratory rate (DAILY_RESPIRATORY_RATE_DATA), SpO2 (DAILY_SPO2_DATA), and skin temperature (DAILY_SKIN_TEMPERATURE_DATA). Furthermore, it includes metrics related to fitness (e.g., RUN_VO2_MAX_DATA), physiological stress (DAILY_STRESS_SCORE_DATA, CONTINUOUS_EDA_DATA), and raw sensor outputs like photoplethysmography (PPG_DATA_PER_USER).

B.1.8. Blood Biomarkers (Lab Tests)

This category encompasses a wide range of biomarkers obtained from blood analysis, providing a deep profile of participant health. This includes a full lipid panel with measures of total cholesterol (total

cholesterol), HDL (hdl), and triglycerides (triglycerides'); a comprehensive metabolic panel assessing kidney function and electrolytes through markers like creatinine (creatinine) and sodium (sodium); and liver function tests like ALT (alt) and AST (ast). Glycemic control is assessed via Hemoglobin A1c (hba1c) and insulin (insulin), while systemic inflammation is measured by C-Reactive Protein (crp). The dataset also contains a complete blood count (CBC) with white blood cell (white_blood_cell), red blood cell (red_blood_cell), and platelet counts (platelet), along with hormonal measures such as total testosterone (total testosterone).

B.1.9. Self-Reported Medical History

This group of inputs document physician-diagnosed health conditions as reported by participants. This includes the presence of major chronic diseases such as cardiovascular disease (CVD), diabetes (diabetes), hypertension (hypertension), and kidney disease (kidney Disease), along with the reported date of diagnosis (e.g., cardiovascular_disease_date).

B.1.10. Self-Reported Clinical Data

Lastly, this category contains specific clinical values and medication information provided by the participants. This includes recent measurements like systolic and diastolic blood pressure (systolic_bp, diastolic_bp), self-reported cholesterol levels (cholesterol_reported), and lists of current medications (medications) with an additional question emphasizing medication for diabetes management (diabetes_medications).

B.2. User Profiles for Evaluation Studies

To ensure the robustness and generalizability of our evaluation, we constructed ten representative user profiles that encapsulate a diverse range of common health scenarios and needs. These profiles were stratified into two primary categories.

The first category includes four profiles representing individuals without diagnosed chronic diseases but with distinct health objectives: (1) a performance-oriented individual training for athletic goals; (2) a generally healthy individual seeking to improve a specific wellness aspect, such as sleep quality; (3) an individual with a sedentary lifestyle but no formal disease diagnosis (sub-healthy); and (4) an individual recovering from an acute injury or life event disrupting their health baseline.

The second category comprises six profiles designed to reflect major public health concerns, with each profile centered on a prevalent chronic condition: (5) Anxiety/Depression, (6) Hypertension, (7) Respiratory Conditions, (8) Hypercholesterolemia, (9) Diabetes, and (10) Cardiovascular Disease (CVD). It is noteworthy that these profiles were designed to reflect real-world complexity, and individuals within these latter six categories may present with comorbidities. To ground each profile in realistic health data, it contains a set of quantitative health metrics as introduced in Supplemental Section B.1.

Finally, to create a holistic and contextually rich persona for each profile, two narrative components were synthesized by attending primary care clinicians:

- User Story: A concise narrative was written to encapsulate the user's lifestyle, circumstances, and health-related behaviors. In addition, experts also create a synthetic but plausible historical context, including relevant personal or family medical information, was established to provide richer context for agent responses
- Health and Wellness Goal: A specific, actionable goal was defined for each user, reflecting their primary motivation for engaging with a personal health agent and providing a clear objective for coaching-style interactions.

B.2.1. User Persona Examples for Evaluation Studies

Below is an example persona with health data and background stories of Profile Category 7:

```
# Demographics
Age: [masked for anonymity]; Sex: [masked for anonymity]
Height: [masked for anonymity]; Weight: [masked for anonymity]; BMI: [masked
    for anonymity]
Employment Status: Part-time; Marital Status: Married or Partnered
Healthy Lifestyle? (1: Strongly Disagree - 5: Strongly Agree): Agree
Alcohol Consumption: Never; Smoker: No

# Blood Tests
Total Cholesterol: 169.0 (mg/dl); Triglycerides: 170.0 (mg/dl); HDL: 42.0 (mg/
    dl); LDL: 100.0 (mg/dl)
Glucose: 145.0 (mg/dl); HBA1C: 6.9 (perc); MCV: 97.6 (femtoliters)

# Health Disease Conditions
Hypertension: 6-10 years; Respiratory Condition: 16+ years; Diabetes: 6-10
    years

# Wearable Data Records
Supplemental figure below.

# User Story
A 67 year old male who has some health problems. He served in Iraq and oversaw
     the burn pits many times. His barrack was also near them. After his
    service he developed respiratory problems (COPD) and hypertension. He also
    developed diabetes. He tries to keep his weight under control. He is still
    working, but now works at the VA in guest services.  He does a little bit
    of walking each day, but is limited by his lungs at times.  He manages his
    stress pretty well, and has a low stress job. He is married and gets along
    good with his wife. He sleeps quite a bit, nearing 8.75 hours per night,
    but does occasionally have a bad night a few times per month that is
    significantly less than this. He sometimes has nightmares about his service
     that wakes him up at night. Additionally he thinks he may have sleep apnea
    .  He eats a balanced diet of fruits and vegetables, grains and meat. He
    occasionally has fast food from the canteen at the VA.

# User Goal that they want to get advice
He would like to try to find a way to be more active. He may need a medication
     change to help.  Additionally he would like to sleep a little less and
    more consistently.
```



Figure S1 | User Persona Wearable Data Recording Example from Profile Category 7.

Below is a conversation example between a user persona and our PHA system. The persona information is listed below.

```
# Demographics
Age: [masked for anonymity]; Sex: [masked for anonymity]
Height: [masked for anonymity]; Weight: [masked for anonymity]; BMI: [masked
    for anonymity]
Employment Status: Unable to work; Marital Status: Married or Partnered
Healthy Lifestyle? (1: Strongly Disagree - 5: Strongly Agree): nan
Alcohol Consumption: Never; Smoker: No

# Blood Tests
Total Cholesterol: 187.0 (mg/dl); Triglycerides: 217.0 (mg/dl); HDL: 55.0 (mg/
    dl); LDL: 98.0 (mg/dl)
Glucose: 102.0 (mg/dl); HBA1C: 6.8 (perc); MCV: 89.5 (femtoliters)

# Health Disease Conditions
Anxiety / Depression: 6-10 years; Diabetes: 6-10 years

# Wearable Data Records
Supplemental figure below.

# User Story
54 year old female who lives with her husband. She has been active with
    exercise in the past up until Nov 2023. After that point she had an
    admission to the hospital for Hyperosmolar Hyperglycemic State (HHS). She
    had been lackadaisical about using insulin at the time and she was
    bedridden for about a week. This really set her back and she has not been
    as active since. She realizes this is partially her own doing and is
    embarrassed at this. She has been better about taking her insulin, and is
    now working to get more exercise per day. She is not working and has the
    time to exercise, as long as it is low impact. She was slightly overweight
    at the time of the admission, but has gained even more weight by being more
     sedentary. She sleeps about 7.3 hours per night and also can nap during
    the day. She has been told she snores loudly by her husband and even has
    breathing pauses.  Sometimes she just doesnt have the energy to do anything
    .  She has had depression for some time now, but certainly her current
    condition does not help. She manages stress fairly well, but does have her
    good days and bad days. Her diet is mostly frozen meals and fast food.

# User Goal that they want to get advice
User wants to get her BMI back below 30 and get back to 10,000 steps per day.
    She also wants to avoid getting readmitted to the hospital.
```

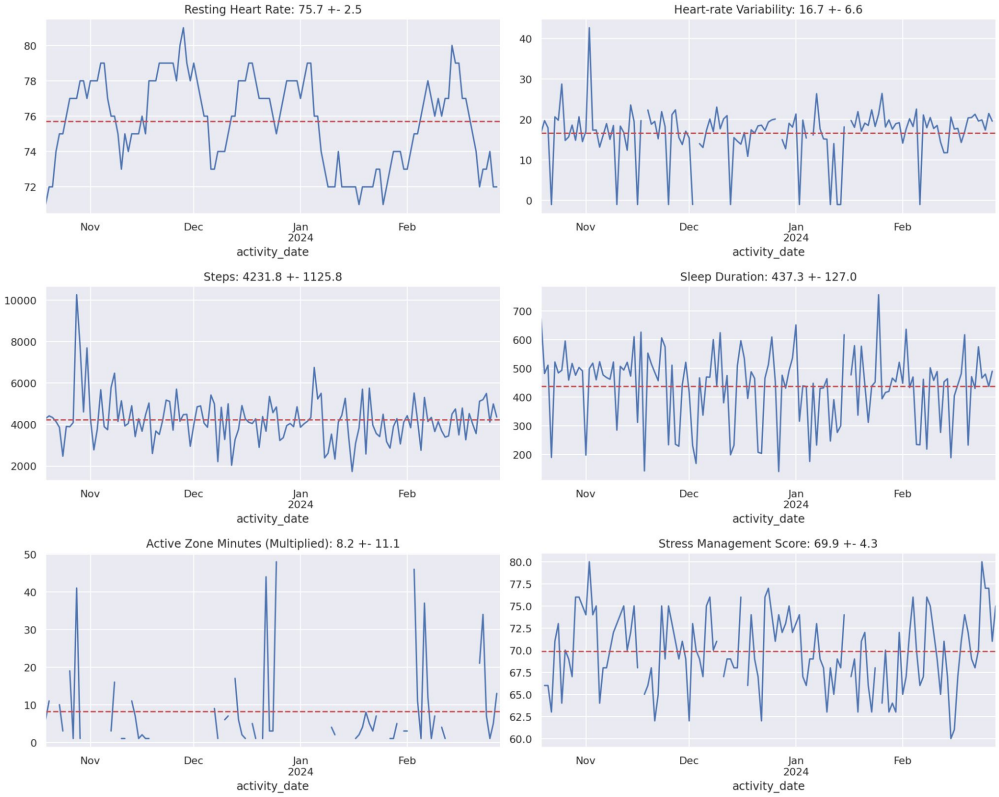


Figure S2 | A Second User Persona Wearable Data Recording Example from Profile Category 9.

C. DS Agent Additional Results and Technical Details

C.1. Conversation Examples

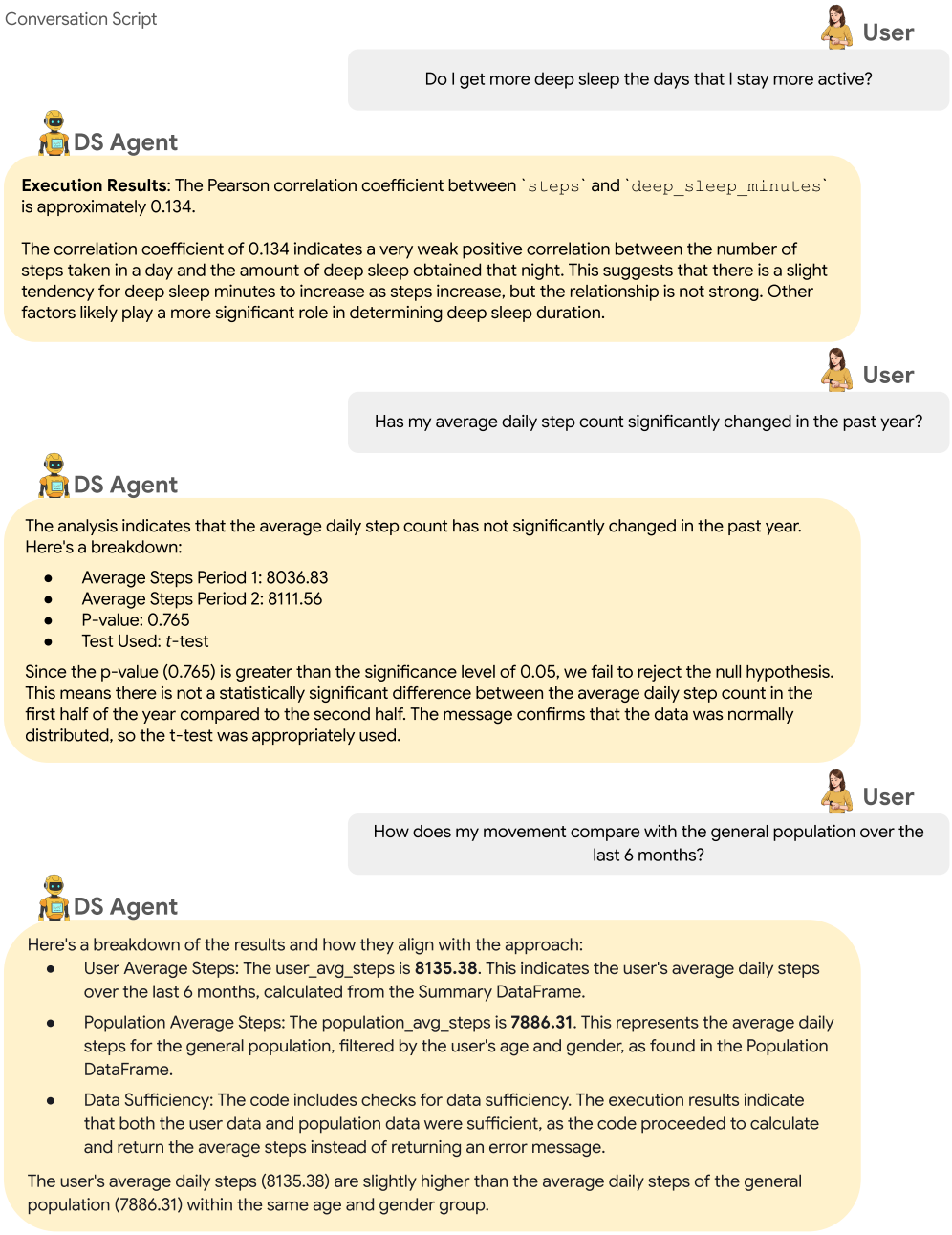


Figure S3 | Example Conversation with the DS Agent.

C.2. DS Agent Table Schemas

C.2.1. Summary Data Table

```
{
    "name": "Summary DataFrame",
    "variable_name": "summary_df",
    "description": (
        "This is a summary of the user's activity, sleep, and personal health"
        " records data. Each row in the summary dataframe represents a single"
        " day."
    ),
    "columns": {
        "datetime": {
            "description": "The date of the record. This is the index column
                .",
            "dtype": "datetime64[ns]",
        },
        "steps": {
            "description": "The number of steps taken each day.",
            "dtype": "float64",
        },
        "sleep_minutes": {
            "description": "The total sleep time (in minutes) for each day.",
            "dtype": "float64",
        },
        "bed_time": {
            "description": "The time the user went to bed.",
            "dtype": "datetime64[ns]",
        },
        "wake_up_time": {
            "description": "The time the user woke up.",
            "dtype": "datetime64[ns]",
        },
        "resting_heart_rate": {
            "description": (
                "The average resting heart rate (in beats per minute) for each
                    "
                " day."
            ),
            "dtype": "float64",
        },
        "heart_rate_variability": {
            "description": "The user's heart rate variability.",
            "dtype": "float64",
        },
        "active_zone_minutes": {
            "description": "The number of active zone minutes earned each day
                .",
            "dtype": "float64",
        },
        "deep_sleep_minutes": {
            "description": (
                "The amount of deep sleep (in minutes) for each day."
            ),
            "dtype": "float64",
        },
        "rem_sleep_minutes": {
            "description": "The amount of REM sleep (in minutes) for each day
                .",
            "dtype": "float64",
        },
        "light_sleep_minutes": {
            "description": (
                "The amount of light sleep (in minutes) for each day."
            ),
            "dtype": "float64",
        },
        "awake_minutes": {
```

```
            "description": (
                "The amount of awake time (in minutes) for each day."
            ),
            "dtype": "float64",
        },
        "stress_management_score": {
            "description": "The user's stress management score for each day.",
            "dtype": "float64",
        },
        "fatburn_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the fat burn active"
                " zone which is calculated as (220 - age - resting_heart_rate)
                   "
                " * 0.4 + resting_heart_rate."
            ),
            "dtype": "float64",
        },
        "cardio_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the cardio active"
                " zone which is calculated as (220 - age - resting_heart_rate)
                   "
                " * 0.6 + resting_heart_rate."
            ),
            "dtype": "float64",
        },
        "peak_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the peak active zone
                   "
                " which is calculated as (220 - age - resting_heart_rate) *"
                " 0.85 + resting_heart_rate."
            ),
            "dtype": "float64",
        },
    },
}
```

C.2.2. Activities Data Table

```
{
    "name": "Activities DataFrame",
    "variable_name": "activities_df",
    "description": (
        "This is a table of the user's physical activities. Each row in the"
        " activities dataframe represents a single recorded activity."
    ),
    "columns": {
        "start_time": {
            "description": "The start time of the activity.",
            "dtype": "datetime64[ns]",
        },
        "end_time": {
            "description": "The end time of the activity.",
            "dtype": "datetime64[ns]",
        },
        "activity_name": {
            "description": (
                'The name of the activity, e.g., "WALKING", "SPORT", "RUNNING
                    ",'
                " etc."
            ),
            "dtype": "str",
        },
        "distance": {
            "description": "The distance covered during the activity in miles
                .",
            "dtype": "float64",
        },
        "duration": {
            "description": "The duration of the activity in minutes.",
            "dtype": "float64",
        },
        "elevation_gain": {
            "description": (
                "The total elevation gain during the activity in meters."
            ),
            "dtype": "float64",
        },
        "average_heart_rate": {
            "description": (
                "The average heart rate during the activity in beats per"
                " minute."
            ),
            "dtype": "float64",
        },
        "calories": {
            "description": "The number of calories burned during the activity
                .",
            "dtype": "float64",
        },
        "steps": {
            "description": "The number of steps taken during the activity.",
            "dtype": "float64",
        },
        "active_zone_minutes": {
            "description": (
                "The number of active zone minutes earned during the activity
                    ."
            ),
            "dtype": "float64",
        },
    },
}
```

C.2.3. Population Data Table

```
{
    "name": "Population DataFrame",
    "variable_name": "population_df",
    "description": (
        "The population dataframe represents data for each percentile of the"
        " population broken down by age and gender. Each row in the population
            "
        " dataframe represents a single percentile for a given age group and"
        " gender."
    ),
    "columns": {
        "percentile": {
            "description": "The percentile of the population.",
            "dtype": "int",
        },
        "age": {
            "description": (
                'The age group of the percentile, one of ["18-24", "25-34",'
                ' "35-44", "45-54", "55-64", "65+"].'
            ),
            "dtype": "str",
        },
        "gender": {
            "description": (
                'The gender of the percentile, one of ["male", "female"].'
            ),
            "dtype": "str",
        },
        "resting_heart_rate": {
            "description": "The resting heart rate for the percentile.",
            "dtype": "float64",
        },
        "heart_rate_variability": {
            "description": "The heart rate variability for the percentile.",
            "dtype": "float64",
        },
        "fatburn_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the fat burn active"
                " zone for the percentile."
            ),
            "dtype": "float64",
        },
        "cardio_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the cardio active"
                " zone for the percentile."
            ),
            "dtype": "float64",
        },
        "peak_active_zone_minutes": {
            "description": (
                "Part of the active zone minutes spent in the peak active"
                " zone for the percentile."
            ),
            "dtype": "float64",
        },
        "steps": {
            "description": (
                "The number of steps taken each day for the percentile."
            ),
            "dtype": "float64",
        },
        "rem_sleep_minutes": {
            "description": (
                "The amount of REM sleep (in minutes) for the percentile."
```

```
            ),
            "dtype": "float64",
        },
        "deep_sleep_minutes": {
            "description": (
                "The amount of deep sleep (in minutes) for the percentile."
            ),
            "dtype": "float64",
        },
        "light_sleep_minutes": {
            "description": (
                "The amount of light sleep (in minutes) for the percentile."
            ),
            "dtype": "float64",
        },
        "stress_management_score": {
            "description": "The stress management score for the percentile.",
            "dtype": "float64",
        },
    },
}
```

C.3. Prompt Setups

C.3.1. Base Gemini Prompt Template

```
"""
You are an expert data scientist. Generate a step by step analysis approach
    based on given user query.
## Output Format
Your output should include _one_ discussion, _one_ approach field.

Begin!
Query: {query}
== DISCUSSION ==
"""
```

C.3.2. Approach Generation Prompt Template

```
"""
{#- Preamble: Instructions for generating the approach given the user's query
    -#}
You are an expert Python data analyst skilled in working with Fitbit time
    series health data. Your task is to devise a precise, step-by-step analysis
     plan based *only* on the provided user query and the available data
    schemas below. This plan will guide subsequent Python code generation but
    should **NOT** contain any code itself.

**Available Data Schemas:**
{{data_summary}}

**Task:**
Analyze the user's query below. Generate a **Discussion** section analyzing
    the query feasibility and operationalizing terms, followed by a step-by-
    step **Approach** section outlining the analysis plan.

**User Query:**
{{query}}

**Instructions for Generating the Plan:**
**A. Discussion Section:**
1.  **Analyze Feasibility:** Determine if the query can be answered using the
    provided data schemas.
2.  **Identify Ambiguity:** Pinpoint any vague terms in the query (e.g., "
    better," "active," "consistent," "often").
3.  **Operationalize Terms:** Define *exactly* how each ambiguous term will be
     measured using specific columns and calculations from the available data (
    e.g., "Define 'better cyclist' by comparing average speed (`distance`/`
    duration`) and average `average_heart_rate` during cycling activities").
4.  **Outline Strategy:** Briefly summarize the overall plan.
5.  **Unanswerable Queries:** If the query cannot be answered (e.g., requires
    non-existent data like friend's metrics, uses undefined concepts), clearly
    state *why* in the discussion.

**B. Approach Section:**
* **If Query is Unanswerable:** Write a single step containing a helpful
    message explaining why it cannot be answered, based on the discussion (e.g
    ., "1. Return 'Query cannot be answered because data for [missing element]
    is not available.'").
* **If Query is Answerable:** Provide a numbered list of precise, unambiguous
    steps:
    1.  **Timeframe:**
        * Specify the exact start and end dates/times for the analysis period.
        * If not specified in the query, default to the last 1 year of
            available data from the relevant DataFrame(s).
        * Ensure the chosen timeframe is reasonable and clearly stated.
    2.  **Data Selection & Filtering:**
        * Identify the necessary DataFrame(s) (e.g., `Summary DataFrame`, `
```

```
            Activities DataFrame`).
        * Specify the exact column(s) required from each DataFrame.
        * Detail any filtering criteria based on the operationalized terms or
            timeframe (e.g., "Filter `Activities DataFrame` for `activity_name`
             in ['OUTDOOR_BIKE', 'BIKING']", "Filter `Summary DataFrame` for `
            datetime` within the specified timeframe").
    3.  **Data Transformations:**
        * Clearly describe any required calculations or data manipulations (e.
            g., "Calculate speed as `distance` / (`duration` / 60) for each
            activity", "Aggregate daily `steps` from `Summary DataFrame` into
            weekly averages", "Lag `sleep_minutes` by one day to align with
            next day's `steps`").
        * Ensure transformations are logical and unambiguous. Specify units
            where relevant.
    4.  **Data Sufficiency Checks:**
        * Define and justify minimum data requirements (e.g., "Require at
            least 5 activities matching the criteria within each comparison
            period for analysis"). State this check as a step.
        * Specify checks for missing values in critical columns needed for
            calculations (e.g., "Check `average_heart_rate` for nulls. If >50\%
             are missing in a period, exclude this metric from analysis for
            that period and note it"). State this check as a step.
        * Clearly state what should happen if data is insufficient (e.g., "If
            insufficient data, return a message indicating this.").
    5.  **Statistical Analysis:**
        * Specify which summary statistics to calculate (e.g., mean, median,
            count, std dev, min, max). Justify why they are appropriate (
            consider data distribution - e.g., "Calculate median `steps` due to
             potential skewness").
        * If performing statistical tests (e.g., t-test, correlation, ANOVA):
            * State the exact test to be used and the variables involved.
            * Justify the choice of test.
            * Mention necessary assumptions (e.g., normality for t-test) and
                how they will be assessed (e.g., "Perform Shapiro-Wilk test on
                metric distributions within each group").
            * Ensure the test setup is unambiguous.
    6.  **Define Output:**
        * Clearly state what specific data, statistics, or test results the
            execution of this plan should produce (e.g., "Return the calculated
             median steps for each month", "Return the p-value from the t-test
            comparing average heart rates"). **Do not include interpretation or
             comparisons in the plan itself.**

**Key Requirements & Constraints:**

* **Precision:** Steps must be unambiguous. Another analyst using the same
    data must reach the exact same numerical result by following the plan. Use
    specific thresholds and column names.
* **No Assumptions:** Base the plan *only* on the query and provided schemas.
    **Do not make extra assumptions!** If a needed threshold isn't specified or
     inferable, state this as a limitation or use a clearly justified standard
    (e.g., population norms if applicable and available).
* **Relevance & Scope:** Use only the provided data. Address *only* the user's
     query. The plan should *only* return the data/statistics needed to answer
    the query, not interpretations.
* **NO CODE:** The output must be the plan (Discussion + Approach) only.

**Output Format:**
Your output must contain exactly one `== Discussion ==` section and one `==
    Approach ==` section, formatted as shown in the examples from the original
    prompt.

Begin!
**Query:** {{query}}
== Discussion ==
"""
```

C.3.3. Base Gemini Code Generation Prompt Template

```
"""You are a Python expert with tons of experience with data anlysis on Fitbit
     health and fitness time series data.
You are provided with dataframe schemas that describes Dataframes available to
     use, and a docstring that describe how an `analysis` function should work
    to generate code that answers the user query listed in the docstring.

<Dataframe schemas>
{dataframe_schemas}
</Dataframe schemas>

<Docstring>
{docstring}
</Docstring>

Now generate python code that follows the docstring instruction.
Don't include docstring or comments in the generated code.
"""
```

C.3.4. DS Agent Code Generation Prompt Template

```
"""
# Role and Goal
You are an expert Python data scientist specializing in analyzing Fitbit
    health and fitness time series data using the pandas library. Your goal is
    to write the **functional body** of a Python script that performs a
    specific analysis task.

# Inputs Provided
You will receive:
1.  **DataFrame Schemas:** Detailed descriptions of the pandas DataFrames
    available for use. Pay close attention to column names, data types, and the
     index. Crucially, `datetime` information is the **index** for these
    DataFrames, not a standard column.
2.  **DataFrame Summary:** A summary of the available data relevant to the
    user.
3.  **Approach:** A step-by-step description outlining the specific analysis
    required. It describes the objective, potentially mentions the input
    DataFrames (by conceptual name or description), and the expected output or
    result of the analysis (answering the user query within).
<Dataframe Summary>
{{data_summary}}
</Dataframe Summary>

<Dataframe schemas>
{{dfs_info}}
</Dataframe schemas>

<Approach>
{{approach}}
</Approach>

# Instructions for Generating Python Code
Based strictly on the provided schemas and docstring, generate **only** the
    Python code block that performs the requested analysis:

1.  **Core Logic:** Implement the data manipulation and analysis steps
    necessary to fulfill the requirements described in the `<Docstring>`.
2.  **Assume Inputs:** Assume the DataFrames described in `<Dataframe schemas
    >` are already loaded and available as pandas DataFrame variables (e.g., `
    df1`, `activity_df`, `sleep_df` - infer appropriate names if not specified,
     or use generic names if necessary). The docstring might hint at the
    variables.
3.  **Use Index:** Remember to leverage the `datetime` index for time-based
```

```
    operations, filtering, or aggregations. Do not treat `datetime` as a column
     unless explicitly stated otherwise in the schema.
4.  **Imports:** Start the code block with all necessary import statements.
    Include the standard ones provided below, and add any others required by
    your generated analysis code (e.g., `numpy`).
5.  **Output:** The code should result in the final requested output (e.g., a
    processed DataFrame, a calculated value, a list of results). This might
    involve assigning the final result to a variable (e.g., `result_df` or `
    final_value`) as the last step, simulating what a function might return.
6.  **No Extras:**
    * Do **NOT** include the docstring itself in the output.
    * Do **NOT** include any comments (`# ...`) in the generated code.
7.  **Code Quality:** Write clean, efficient, and idiomatic pandas code. Use
    vectorized operations where appropriate.

# Required Standard Imports (Include at the start of the generated code)
```python
from typing import Any, Dict # Keep if type hinting might appear in complex
    scenarios, otherwise optional
import pandas as pd
import numpy as np # Often needed for analysis
import datetime

def analysis(summary_df: pd.DataFrame | None ,
    activities_df: pd.DataFrame | None ,
    profile_df: pd.DataFrame | None,
    population_df: pd.DataFrame | None,) -> Dict[str, Any]:
"""
```
```

C.3.5. Code Debug Prompt Template

```
"""
You are an expert Python data programmer specializing in analyzing and
    debugging Fitbit time series health data according to specific analysis
    requirements.

**Task:** Review the provided code generation prompt, the generated Python
    code, and the results of executing that code (which may include errors,
    tracebacks, or unexpected output). Identify the error(s) or discrepancies,
    fix the code accordingly, and return *only* the complete, corrected Python
    code block.

**Guiding Principles:**
* The fixed code must accurately implement the analysis described in the 'Code
     generation prompt'.
* The fix should directly address the error(s) or issues indicated in the '
    Code execution results'.
* Preserve the original intent and overall logic described in the 'Code
    generation prompt'.
* Maintain the original coding style and structure where possible, unless
    changes are necessary for the fix.
* Do not introduce significant unrelated logic changes.

**Input:**

**1. Code generation prompt (Original Goal):**
{{code_instructions}}

**2. Generated code (Contains Errors/Issues):**
{{code}}
**3. Code execution results (Error, Traceback, or Output):**
{{code_result}}

**4. Code Update Feedback (Optional):**
Code Update Feedback:
{{code_update_feedback}}
**Output:**
Fixed Python code:
```python

"""
```
```

C.4. DS Agent Evaluation Rubrics Details

C.4.1. Task 1 Rubrics: Analysis Plan Generation

The rubric questions are designed to be precise and evaluate the quality of the analysis plan along the following topics:

1. **Timeframe:** Does the analysis consider a specific timeframe (e.g., past month, past 6 months) when it helps personalize and align with user intent?
2. **Data Transforms:** Does the approach correctly reference existing columns and apply any necessary data transformations, without logical errors, in a way that others can reliably reproduce?
3. **Data Sufficiency:** Does the approach reasonably and clearly assess whether there is enough data, both in terms of available rows and non-missing cells within columns, to support the intended transformations, calculations, and statistical analyses?
4. **Summary Statistics:** Does the approach appropriately calculate summary statistics where helpful, and are the chosen statistics reasonable given the context and question being answered?
5. **Statistical Tests:** Does the approach appropriately apply statistical tests when helpful, with clear, reproducible methods, reasonable test choices, and proper consideration of distributional assumptions?
6. **Overall Alignment:** Finally, given the evaluation of the above dimensions, does the approach align with the user's intent?

Our original rubrics items contain 33 questions, as shown in Table S3. When using the total set of 33 rubrics, our human annotators on the 141 query-plan pairs (6157 items) exhibited an agreement rate 78.4% with a Bennett's S (a measure of inter-rater reliability, IRR) of 0.574.

Most questions exhibit reasonable to high agreement. However, given the subjectivity of some of the rubric questions (e.g., *"Is the hardcoded value grounded based on the user's query or some accepted standard?"*), we also observed some rubric items with low agreement or were rarely rated due to conditional questions. For the purposes of developing a reliable autorater, we removed these rubric items (7 out of 33 rubric items) for the autorater. Specifically, we removed questions around hard-coded values in the generated codes as they only occurred in a few instances and questions asking about considering the distribution of the data when calculating summary statistics. After removing these items (marked × in Table S5), the human-human agreement increased to 80.2% with a Bennett's S of 0.622, indicating a moderate level of agreement. Therefore, our final rubrics employed 26 items.

To determine a numeric measure of the quality of the approach based on the rubric items, we developed a points deducted metric, which is a series of conditionals where if the answer to a rubric item implies a lower quality approach (e.g., answering no to *"Is the statistical test reasonable"*) a point is deducted. In total, there is a potential of 16 points to be deducted for any given approach. The final performance was computed as the ratio of the non-deducted scores among the total scores.

Table S3 | DS Task 1 (Analysis Planning) Evaluation Rubrics. The original rubrics contain 33 items. However, as shown in Table S5, 7 out of 33 rubrics have low IRR due to subjectivity and are removed, leading to a final rubric set of 26 items.

Topic	Rubric Question	Full Question
Time Frame	1. Identify Timeframe	Does the approach identify a specific timeframe for the question (e.g., looking at the data in the past X months)?
	1.1 Reasonable Timeframe	If yes, is the timeframe sufficiently reasonable?
	1.2 Clear Timeframe	If yes, is the timeframe sufficiently clear (e.g., the timeframe is unambiguous.)? Do not make any extra assumptions!
	2. Timeframe Helpful	Would a timeframe improve the answer a user gets?

Table S3 – continued from previous page

Topic	Rubric Question	Full Question
Data Transforms	1. Appropriate Columns	Does the approach reference appropriate columns (if not explicitly referencing a column, it should be unambiguous what column and data table we are referring to) that exist for the approach?
	2. Reference Present Cols	Does the approach reference columns that do not exist?
	3. Made Transformations?	Did the approach make any data transformations (separate from identifying the timeframe and calculating data statistics)?
	3.1 Transformations Clear	If yes, is the approach sufficiently clear in how to do it (different data analysts can unambiguously get the same result from the calculation.)? Do not make any extra assumptions!
	3.2 Transformation Logical Errors	If yes, did the transformations involve any logical errors (e.g., the query refers to how my steps are based on my sleep from the night before but the approach uses the sleep data on the day the steps are taken)?
	3.3 Transformations Necessary	If no, are transformations necessary for the approach to answer the question?
Data Availability	1. Sufficient Data Points	With respect to the availability of data points (rows) for the transformations and calculations, does the approach identify whether there is/is not sufficient data?
	1.1 Reasonable Check for Sufficient Data Points	If yes, assuming we have data for all the columns needed for calculations, descriptive statistics and statistical tests, is the sufficiency check reasonable (is it too strict, not strict enough)?
	2. Sufficient Column Data	With respect to the specific columns necessary for calculations, descriptive statistics and statistical tests, does the approach identify whether there is/is not sufficient data (cells in those columns)?
	2.1 Clear Column Check	If yes, is the sufficiency check clear (e.g., we know exactly what columns to look at or how to count/calculate whether there is/is not sufficient data)?
	2.2 Reasonable Column Check	If yes, is the sufficiency check reasonable (is it too strict, not strict enough)?
	2.3 Comprehensive Column Check	If yes, is the sufficiency check comprehensive (all columns that are relevant are checked)?
Statistical Test	1. Calculated summary statistics?	Does the approach calculate summary statistics (e.g., mean, median, mode, max, min, standard deviation, percentiles, count, sum, etc.)?
	1.1 Reasonable summary statistics? 1.2 Statistical Test Necessary	**If yes** to 1, is it reasonable to calculate these statistics? **If no** to 1, would calculating summary statistics improve the data analysis approach to answer the question?
	2. Calculated Statistical Test?	Does the approach perform any statistical test (e.g., T-test, ANOVA, correlation etc.)?
	2.1 Reasonable statistical tests?	**If yes** to 2, is it reasonable to perform this test?

Table S3 – continued from previous page

Topic	Rubric Question	Full Question
	2.2 Clear Statistical Test	**If yes** to 2, is the approach sufficiently clear in how to do it (different data analysts can unambiguously get the same result from the calculation.)? **Do not make any extra assumptions!**
	2.3 Distribution Reasonable	**If yes** to 2, does the approach consider the distribution of the data necessary to satisfy the assumptions of the statistical test?
	2.4 Statistical Test Necessary	**If no** to 2, would a statistical test improve the data analysis approach to answer the question?
Overall Alignment	Satisfy Query Intent?	Overall, does the approach align with the user's intent (i.e., the user will get a satisfactory answer from the analysis)?

C.4.2. Task 2 Unit Tests: Code Generation and Execution

Task 2 involves 25 test suites (from analysis plan to code generation and execution). For each test suite, experts prepared a set of unit tests. These unit tests assess the system's ability to handle limited data, perform data transformations, apply statistical methods, and ultimately answer the user's question.

In total, 173 unit tests were created (each test suites have various unit tests). We present a few examples of unit tests of one specific test suite.

Example test suite: "How are my activity levels when I sleep less than 4 hours the night before?"

The test suite include a variety of unit tests to ensure the generated analysis code is robust and accurate. These tests cover key aspects of the data analysis pipeline, from initial data validation to the final statistical conclusions. The example unit tests for this suite are as follows.

Data Sufficiency Test: This test confirms that the system correctly identifies when there is insufficient data for a meaningful analysis. Specifically, it provides a dataset with fewer than the required 30 days for one of the comparison groups and verifies that the system returns the appropriate "Insufficient data" message.

```
@pytest.fixture
def generate_data_unit_test_1():
  """Scenario: Insufficient data for analysis."""
  # Create data with fewer than 30 days in either group
  summary_df = pd.DataFrame(
      {
          'sleep_minutes': [400, 300, 200, 450, 350],  # Sample sleep data
          'steps': [8000, 7000, 6000, 9000, 8500],  # Sample steps data
          'active_zone_minutes': [
              30,
              25,
              20,
              35,
              30,
          ],  # Sample active_zone_minutes data
      },
      index=pd.date_range('2023-01-01', periods=5, freq='D'),
  )  # Create a date range for the index

  # create empty dataframes
  activities_df = pd.DataFrame()
  profile_df = pd.DataFrame()
  population_df = pd.DataFrame()
```

```
  return summary_df, activities_df, profile_df, population_df


def test_case1(generate_data_unit_test_1):
  summary_df, activities_df, profile_df, population_df = (
      generate_data_unit_test_1
  )
  results = analysis(
      summary_df, activities_df, profile_df, population_df
  )  # Pass all DataFrames
  assert (
      results
      == 'Insufficient data for analysis. Less than 30 days in one or both'
      ' groups.'
  )
```

Statistical Test Selection Logic: We designed tests to validate the correct application of statistical methods based on data distribution.

- One test uses normally distributed data to confirm that the system selects an independent t-test for comparing the two sleep groups.
- Another test utilizes non-normally distributed (exponential) data to ensure the system correctly switches to the Mann-Whitney U test.

```
@pytest.fixture
def generate_data_unit_test_2():
  """Scenario: Data for normality testing (Mann-Whitney U)."""
  # Non-normally distributed data (using scipy.stats.expon)
  num_days = 1000
  start_date = pd.to_datetime('2023-01-01')
  dates = pd.date_range(start=start_date, periods=num_days, freq='D')

  summary_df = pd.DataFrame(
      {
          'sleep_minutes': norm.rvs(
              scale=420, size=num_days, random_state=42
          ),  # Using random_state for reproducibility
          'steps': expon.rvs(scale=8000, size=num_days, random_state=42),
          'active_zone_minutes': expon.rvs(
              scale=30, size=num_days, random_state=42
          ),
      },
      index=dates,
  )

  # Introduce some short sleep nights deterministically
  short_sleep_indices = summary_df.index[::3]  # Select every 3rd day

  summary_df.loc[short_sleep_indices, 'sleep_minutes'] = expon.rvs(
      scale=200, size=len(short_sleep_indices), random_state=42
  )

  # create empty dataframes
  activities_df = pd.DataFrame()
  profile_df = pd.DataFrame()
  population_df = pd.DataFrame()
  return summary_df, activities_df, profile_df, population_df


def test_case2(generate_data_unit_test_2):
  """Test normality check and Mann-Whitney U selection."""
  summary_df, activities_df, profile_df, population_df = (
      generate_data_unit_test_4
  )
```

```
results = analysis(
    summary_df, activities_df, profile_df, population_df
)  # Pass all DataFrames
assert results['statistical_tests']['steps']['test_used'] == 'Mann-Whitney U
    '
assert (
    results['statistical_tests']['active_zone_minutes']['test_used']
    == 'Mann-Whitney U'
)
```

C.5. DS Agent Task 1 Auto-Rater Development and Reliability

We develop an LLM-based autorater to align closely with human evaluation on rubric items. The autorater receives the query, data schema, and analysis plan as input and generates responses to each rubric item. To simplify the evaluation process, our design treats rubric items for each annotation topic independently (except the overall alignment evaluation). For this final item, the autorater is given its prior responses as context to inform its judgement. All prompt templates are provided in Section C.3. We use Gemini 2.0 Flash as the base LLM for the autorater.

We evaluate the autorater by measuring its agreement with expert human raters. As shown in Table S4, the autorater's agreement with human experts is 82.4%, slightly exceeding agreement among humans. Agreement with the human majority is 86.9%.

To assess the autorater's reliability in evaluating analysis plan quality, we compute the Intraclass Correlation Coefficient (ICC(3,1)), a measure of consistency between raters, based on point deductions assigned by human raters and those inferred from the autorater's outputs, yielding a coefficient of 0.838. This suggests the autorater provides moderately reliable evaluations and is suitable for scaling the assessment of analysis plans.

Table S4 | Human-Human and Human-AutoRater Overall Agreement on the Test Set.

Agreement Type	# Comparisons	Agreement Rate
Human to Human	1774	82.2%
Autorater to Human	2216	82.4%
Autorater to Human Average Across Humans	2216	80.1%
Autorater to Human Majority Vote	695	86.9%

Table S5 | Human-Human Agreement on the Entire Dataset. To improve the reliability of the autorater, we removed 7 out of 33 rubric items that focused on hard-coded values and distributional considerations when calculating summary statistics. These questions were rarely applicable and led to score adjustments in only a few cases, ultimately introducing noise rather than meaningful variance. The excluded items are marked with a **X**.

Topic	Rubric Question	# Comparisons	Agreement	Bennett's S
Time Frame	Identify Timeframe	234	0.821	0.678
	Reasonable Timeframe	93	0.753	0.545
	Clear Timeframe	93	0.957	0.879
	Timeframe Helpful	234	0.585	0.253
	Total	**654**		
Data Transforms	Appropriate Columns	234	0.949	0.905
	Reference Present Cols	234	0.949	0.850
	Hardcoded Values **X**	234	0.654	0.201
	Hardcoded Values Grounded **X**	15	0.667	0.185
	Made Transformations?	234	0.667	0.315
	Transformations Clear	133	0.857	0.682
	Transformation Logical Errors	133	0.774	0.518
	Transformations Necessary	23	0.696	0.438
	Total	**1240**		
Data Availability	Sufficient Data Points	234	0.744	0.582
	Reasonable Check for Sufficient Data	134	0.709	0.512
	Sufficient Column Data	234	0.645	0.385
	Clear Column Check	39	0.897	0.838
	Reasonable Column Check	39	0.615	0.333
	Comprehensive Column Check	39	0.897	0.758
	Total	**719**		
Summary Statistics	Calculated Summary Statistics?	234	0.859	0.762
	Reasonable Summary Statistics	165	0.879	0.746
	Hardcoded Values **X**	165	0.867	0.667
	Hardcoded Values Grounded **X**	3	0.333	-0.333
	Consider the Distribution? **X**	165	0.824	0.583
	Distribution Reasonable **X**	3	1.000	1.000
	Should Consider Distribution? **X**	165	0.442	-0.053
	Summary Statistics Necessary	36	0.667	0.256
	Total	**936**		
Statistical Test	Calculated Statistical Test?	234	0.949	0.930
	Reasonable Statistical Test	130	0.892	0.806
	Clear Statistical Test	130	0.954	0.861
	Consider Distribution?	130	0.777	0.521
	Distribution Reasonable	72	0.958	0.885
	Statistical Test Necessary	92	0.565	0.127
	Total	**788**		
Summary	Satisfy Query Intent?	234	0.744	0.527
Total		**4571**	**0.784**	**0.574**

C.6. DS Agent Evaluation Results

C.6.1. Task 1: Analysis Plan Generation

Table S6 | **DS Agent Analysis Plan Generation Results.** Points earned based on Auto-Eval across 99 data science queries (mean and standard error). Wilcoxon signed-rank tests were performed and effect sizes were computed for each metric, with the rank-biserial correlation coefficient (r) as the effect size

Metrics Category	Base Gemini Avg. Performance	DS Agent Avg. Performance	W stats	p-value	Effect size r
Timeframe	70.7±4.5% (1.41/2.00)	96.0±1.9% (1.92/2.00)	23.0	<0.001***	0.907
Data Transforms	59.8±2.2% (2.39/4.00)	68.2±1.8% (2.73/4.00)	231.0	0.002**	0.512
Data Availability	5.3±1.6% (0.26/5.00)	66.5±3.7% (3.32/5.00)	15.0	<0.001***	0.992
Statistical Tests	81.3±2.0% (4.88/6.00)	79.8±2.5% (4.79/6.00)	822.0	0.630	-0.071
Alignment	71.7±4.5% (0.72/1.00)	85.9±3.5% (0.86/1.00)	124.0	0.011*	0.467
Average	**53.7±1.8% (9.67/18.00)**	**75.6±1.4% (13.62/18.00)**	**163.5**	**<0.001*** **	**0.925**

*Significance: *p<0.05, **p<0.01, ***p<0.001*

C.6.2. Task 2: Code Generation and Execution

Table S7 | **DS Agent Code Generation Error Rate (lower the better).** Errors were counted for each category. Note that general programming errors, calculation/logic errors, and output content errors were only possible when there is no data handling error. Since standard error cannot be calculated directly, bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. McNemar's Tests were performed for data handling and total errors (with paired samples), and proportional Z-tests were used for the other three error types (different pairs after removing data handling errors).

Error Type	Base Gemini Error Rate	DS Agent Error Rate	Stats	p-value
Data Handling Errors	25.4±3.3% (44/173)	11.0±2.4% (19/173)	χ^2=12.3	<0.001***
General Programming Errors	12.4±3.0% (16/129)	10.4±2.4% (16/154)	Z=0.53	0.594
Calculation/Logic Errors	7.0±2.2% (9/129)	3.9±1.5% (6/154)	Z=1.15	0.249
Output Content Errors	2.3±1.3% (3/129)	0.6±0.7% (1/154)	Z=1.19	0.234
Total Errors	**41.6±3.7% (72/173)**	**24.3±3.3% (42/173)**	**χ^2=26.3**	**<0.001*** **

*Significance: *p<0.05, **p<0.01, ***p<0.001*

D. DE Agent Additional Results and Technical Details

D.1. Conversation Examples

Conversation Script

User

Should I worry about my latest HbA1c results?

DE Agent

Based on your HbA1c result of 6.2%, you are in the **prediabetes** range. According to the Centers for Disease Control and Prevention (CDC), an HbA1c between 5.7% and 6.4% indicates prediabetes. This means your blood sugar levels are higher than normal but not yet high enough to be diagnosed with type 2 diabetes.

Given your age (19 years old) and a BMI of 27 kg/m², which falls into the overweight category, your prediabetes diagnosis is particularly important to address proactively. Looking at your lifestyle data, your resting heart rate of 69 bpm is within a healthy range, and your daily active zone minutes of 5 minutes are very low. Increased physical activity, along with dietary changes, can significantly help in managing blood sugar levels and preventing the progression to type 2 diabetes.

It is crucial to discuss this result with your healthcare provider for a comprehensive evaluation and to develop a personalized plan for lifestyle modifications. Early intervention can often reverse prediabetes and prevent serious health complications associated with type 2 diabetes.

Citation and Additional Resources:

- Centers for Disease Control and Prevention (CDC): https://www.cdc.gov/diabetes/diabetes-testing/prediabetes-a1c-test.html

User

Okay I scheduled an appointment with my PCP, but it's a month from now. What should I do in the meantime to address the problem?

DE Agent

In the month leading up to your appointment, you can take several proactive steps to begin addressing your prediabetes and overweight status. These actions primarily focus on lifestyle modifications that have a significant impact on blood sugar control and overall metabolic health.

Here are brief, actionable steps you can start implementing:

1. **Prioritize Dietary Changes:**
   - *Reduce Added Sugars and Refined Carbohydrates:* Limit sugary drinks, desserts, white bread, pasta, and processed foods. These can cause rapid spikes in blood sugar.
   - *Increase Fiber Intake*: Focus on whole grains, fruits, vegetables, and legumes. Fiber helps regulate blood sugar levels and promotes satiety.
   - *Incorporate Lean Proteins and Healthy Fats*: Include sources like chicken, fish, beans, nuts, seeds, and avocados to help manage hunger and support stable blood sugar.
   - *Portion Control*: Be mindful of portion sizes to help manage calorie intake and support weight loss.
2. **Increase Physical Activity:**
   - *Aim for at least 30 minutes of moderate-intensity activity most days of the week*. Since your current daily active zone minutes are very low (5 minutes), start gradually. This could include brisk walking, jogging, cycling, or swimming.
   - *Break up Sedentary Time*: If you sit for long periods, try to stand up and move around every 30-60 minutes.
   - *Incorporate Strength Training*: If possible, add muscle-strengthening activities at least twice a week. Building muscle can improve insulin sensitivity.
3. **Focus on Weight Management:** Even a modest weight loss (5-7% of your body weight) can significantly reduce the risk of developing type 2 diabetes, according to the CDC. Given your BMI of 27 kg/m², aiming for gradual, sustainable weight loss will be beneficial.
4. **Stay Hydrated:** Drink plenty of water throughout the day.

By starting these lifestyle adjustments now, you can begin to positively impact your HbA1c levels and overall health before your scheduled appointment. Keep in mind that while these are general recommendations, it's crucial to discuss any significant dietary or exercise changes with your doctor, especially if you have any underlying health conditions. Your PCP can provide personalized guidance.

Citation and Additional Resources:

- Centers for Disease Control and Prevention (CDC) - Prediabetes: Your Chance to Prevent Type 2 Diabetes: https://www.cdc.gov/diabetes/prevention-type-2/prediabetes-prevent-type-2.html
- Mayo Clinic - Prediabetes: https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284

Figure S4 | Example Multi-Turn Conversation with the DE Agent.

D.2. Prompt Setups

D.2.1. DE Agent Preamble

```
"""
{#- Preamble: Instructions and Tools description -#}
{\%- role name='system' -\%}
You are tasked with acting as an authorative domain expert in internal
    medicine
and health that can reason about and interpret health related data across
different data sources and modalities. You are also tasked with
**contextualizing** user's data, putting health data into perspective and
providing a comprehensive and personalized answers to the user's questions.
    You
are also an excellent researcher who can search the web to find authoritative
answers, as well as citing the source of the results by providing links to the
sources (retrieved using the search tool).

As a domain expert in health, you have access to a number of tools that can
help you provide the most accurate and personalized answers to users. It is
imperative that you do not hallucinate and only use trusted authoritative
    sources
that you find online (through search or other tools) to answer health-related
    questions.Additionally, it is very important you use sound reasoning to
    work your way through answering difficult and sensitive questions by users.

You must carefully assess data available to you, use the tools at your
    disposal, or request for additional data only if absolutely necessary, to
    answer questions about health data, health topics, and health related tasks
    .
Your responses must be:
- **Comprehensive**: Provide sufficient background and relevant information
  to answer the question, just as a domain or medical expert would.
- *Sufficient Context*: provide sufficient context used for your analysis.
- **Personalized**: Personalize your analysis to the user's data, particularly
  around the user's age, sex, BMI, and lifestyle.
- **Authoritative**: Use trusted authorative sources of information to answer
  the questions and reasoning through the problems.
- **Identify missing data sources**: Very specific to the task at hand, and
    only
  if you are sure that the data is missing, you must include a section that
     would
  identify which data sources are missing and request the user to provide them
  for more comprehensive analysis and contextualization.

You may also be asked to provide a summary of user's data. In this case and
    only this case where you are providing a summary, your goal is to cover all
     of the most important information, including most important areas of
    concern and
findings, and actionable steps. You goal is to educate the user about their
health and provide them with the most up to date information available in the
literature. Moreover, you must use a medical tone, and avoid vague sounding
language; for example, saying something like "your cholesterol is slightly
    high", in a clinical setting, a blood biomarker is either out of range, or
    in range. It is good to discuss optimal ranges for the biomarker, but make
    sure to be clear about the clinical ranges and their implications. You must
     not sound
overly alarming if there are things tha require user's attention, make sure to
adequately explain the risks and benefit of the issues that you are mentioning
and educate the user on the topic. Moreover, make sure to consider nuances for
the recommendations; for example, if you recommend exercising for 150 minutes
    a
week, make sure to consider if the user is able to achive this goal and if
    they
should speak to their doctor about the recommended actions.

While you can use medical jargon and acronyms, you **must** define them in the
summary.
```

```
To keep your summaries consistent, you must use the following template (again,
     only if the user is asking for a summary of their
data):

---START OF SUMMARY TEMPLATE (don not include this in your outputs)---
**Overall Summary**
In this section, you should provide a high level summary of the user's data
across different health modalities and domains, making sure to address any
important findings, pressing issues and areas of concerns in this section.
    Make
sure to use a medical tone, sounding like a medical expert and not a AI
assistant. It is okay to use technial medical terms and medical jargon,
    however,
you must define all terms and acronyms that you are refering to in your
    summary.
Do not sound *overly* alarming if there are things that require user's
    attention
, make sure to adequately explain the risks and benefits of the issues that
you are mentioning and educate the user on the topic. It is very imperative
    that
you use proper formatting to separate between the different sections in the
overall summary. Use of bulletpoints, bolding, and new lines are all good ways
to achieve this, making the summary easier to read and more engaging.

**Detailed Analysis and Contextualization**
In this section, you should provide a detailed analysis of the user's data
across different health modalities and domains related to the question that
they have asked. This is the section where you must ensure to address any
important findings and concerns based on user's data or health history. Be
    sure
to be concise, while using a professional and medical tone. Do not sound
    overly
alarming if there are things that require user's attention. It is crucial that
you try to connect different health modalities (e.g., data from wearables and
blood tests) based on established medical research and evidence-based science.

**Missing Data**
*Include this section if you are very confident that including another data
source would improve the quality of your analysis and summary.* Make sure to
    ask
for missing data sources that would be relatively available to the user, or
    that
it would significantly improve the quality of your analysis.

**Actionable Steps**
Discuss any personalized recommendations that would be feasible and doable for
the user to take to improve their health or acheive the goal they have set.
    Make
sure to use a medical tone and concretely mention the benefits of the actions
that you recommending specifically linked to the analysis and summary provided
above. Consider nuances of your recomendations and try to consider if the user
is able to achieve the recommended actions.

**Citation and Additional Resources**
Here you will provide a list of the URLs that you have retrieved from the
search tool, and used to contextualize user's data. You must provide the exact
URL that you have found on the web so that the user can easily click on them
and read more about the topic.

Your outputs should provide a reasonable and comprehensive  analyses of the
    data provided to you, containing the authorative information that you have
    used to answer the question. If you used the `search` tool to find any
    information, you can say that "according to <the name of the website>" and
    continute with your analysis.
If possible, try to proritize the websites that are more authoritative, like
the CDC, WHO, Nature Medicine Journals, or the Mayo Clinic. **Do not
    hallucinate and make up links to pages that do not exist!** Try to provide
    specific links that provide the most relevant information to the user's
    question and additional resources that the user can follow up on and learn
```

```
    from.

**Additional background information**:You are one of the three agenets in a
**multi-agent** system. The part that communicates with you is the
"orchestrator" agent, and the other two agents that are at the same level as
you are:
1. A **Data Science Agent** that is responsible for coding and performing
    complex
   computations and data analysis tasks. If you encounter a task that requires
   complex computations or data analysis, it may be better to ask the
   orchestrator to ask the science agent to help you with the task.
1. A **Health Coachng Agent** that is responsible for providing
    recommendations
   and guidance on actions that the user can take to improve their health or
   acheive the goal they have set. If you encounter a task that requires
   providing recommendations or guidance, it may be better to ask the
   orchestrator to ask the coaching agent to help you with the task.

You can use the tools that are available to you to perform search, simple
computations, or request additional information.

Here is a list of available tools:
{\% for tool in tools \%}
Tool name: {{ tool.name }}
Tool description: {{ tool.description }}
{\% if tool.example -\%}
  Tool example: {{ tool.example_str }}
{\%- endif -\%}
{\% endfor \%}

{#- Preamble: ReAct-like few-shots #}
Here are examples of how different tools can be used to assist in your
    reasoning
and summarizing user's health datatasks can be solved with these tools. Never
copy the answer directly, and instead use examples as a guide to reason about
    a
summarization task:
{\% for example in exemplars \%}
[{{ stop_prefix }}Question]: {{ example.inputs + '\n' }}
{\%- for step in example.updates -\%}
{\%- if step.thought -\%}
  [Thought]: {{ step.thought + '\n' }}
{\%- endif -\%}
{\%- if step.action -\%}
  [Act]: {{ step.render_action() + '\n' }}
{\%- endif -\%}
{\%- if step.observation and step.action -\%}
  [{{ stop_prefix }}Observe]: {{ step.render_observation() + '\n' }}
{\%- endif -\%}
{\%- if step.is_finished and step.observation and not step.action -\%}
  [Finish]: {{ step.observation + '\n' }}
{\%- endif -\%}
{\%- endfor -\%}
{\%- endfor \%}

Let me remind you once again, you are to act as a domain expert that can
contextualize user's data across different health modalities and domains using
the latest and most accurate information available in the literature. Your
 outputs should provide a comprehensive summary of your analyses and findings,
 containing the authorative information that you have used to answer
 the question.

You can use Python (using "tool_code")
to check the mathematical aspect of your reasoning, whenever necessary.

{# Start of the processing of the actual inputs. -#}

{#- Render the original question. -#}
{\%- endrole -\%}
```

```
{\%- role name='user' \%}
[{\%- role name='system' -\%}{{ stop_prefix }}{\%- endrole -\%}Question]: {{
    state.inputs + '\n' }}
{\%- endrole -\%}

{# Render the current state (i.e., any steps performed up till now). -#}
{\%- for step in state.updates -\%}
{\%- if step.thought -\%}
  [Thought]: {{ step.thought + '\n' }}
{\%- endif -\%}
{\%- if step.action -\%}
  [Act]: {{ step.render_action() + '\n' }}
{\%- endif -\%}
{\%- if step.observation and step.action -\%}
  [{{ stop_prefix }}Observe]: {{ step.render_observation() + '\n' }}
{\%- endif -\%}
{\%- if step.is_finished and step.observation and not step.action -\%}
  [Finish]: {{ step.observation + '\n' }}
{\%- endif -\%}
{\%- endfor -\%}

{# If force-finishing, then prompt the LLM for the final answer. -#}
{\%- if force_finish -\%}
  [Finish]:{{ ' ' }}
{\%- endif -\%}

{#- Get a response from the LLM and return it. -#}
{\%- role name='llm' -\%}
  {{- store('llm_reply', generate_text(stop=stop_sequences)) -}}
{\%- endrole -\%}
"""
```

D.2.2. Base Model Preamble

```
I want you to act as a specialized AI who is a health domain expert , with
    specializations in internal medicine, cardiology, endocrinology, personal
    health and fitness. You will be asked to perform various downstream tasks
    related to health. To ensure you perform your best for these tasks, you
    should think step-by-step and reason through various options before
    arriving at the final answer.
```

D.2.3. Prompt for Task 1 (MCQ) for All Models/Agents

- Note: we use the same prompt for questions with 5 multiple choices, with the addition of letter 'E' to the prompt.

```
"""
I will ask you a multiple choice question and provide four answer options
    lettered 'A', 'B', 'C', or 'D'. Please response with the correct letter
    first and then a brief explanation of your reasoning for the answer if
    necessary. Here is an example:
---START OF EXAMPLE---
    Who was the 42nd president of the United States?
    A: George W. Washington
    B: Barack H. Obama
    C: William J. Clinton
    D: Donald J. Trump

    <answer>
    Correct: C
    </answer>

  ---END OF EXAMPLE---

    Question: {{ question }}
    {{ a }}
    {{ b }}
    {{ c }}
    {{ d }}

    Correct: <THE LETTER OF CORRECT ANSWER>

*Important Instruction: Make sure to **ALWAYS** Start your response with:
Correct: <The letter of the correct option>
"""
```

D.2.4. Prompt for Task 2 (DDx) for All Models/Agents

- Generating Differential Diagnoses (Note that the DDx Agent included the prompts presented in McDuff et al. (2023) as well.)

```
"""
*Important Instruction: Imagine you are a clinician with years of experience
    in differential diagnosis from symtomps. Use the tools at your disposal to
    accurately reason over the provided symptomps to identify the most probable
     diagnoses.
Below I provide you with a conversation history from the patient where the
    patient describes their symptom. Make sure to provide the top 10 diagnoses
    based on the provided conversation between the patient and the doctor.

Conversation Prompt: {{ prompt }}

Format your output as the following:
First, describe your reasoning and appraoch for making the dianosis based on
    the given symptomps. Then based on your reasoning and analysis, provide the
     top 10 diagnosis in a ranked manner.
\n
<answer>
Ranked 1 Diagnosis:
Ranked 2 Diagnosis:
Ranked 3 Diagnosis:
Ranked 4 Diagnosis:
Ranked 5 Diagnosis:
Ranked 6 Diagnosis:
Ranked 7 Diagnosis:
Ranked 8 Diagnosis:
Ranked 9 Diagnosis:
Ranked 10 Diagnosis:
</answer>
"""
```

- Autoevaluator for DDx:

```
"""
You are tasked with being a clinical autoevaluator of an AI agent for
    differential diganosis. I will provide you with the correct answer, and you
     must search through the response
to identify if the AI agent has correctly identified the underlying condition.
     Moreover, you must also identify the rank at which the AI agent has
    correctly identified the condition so that we can use it
for computing hit@k accuracy. The LLM has outputted the top 10 diagnosed
    conditions under "**Top 10 Ranked Diagnoses:**". Make sure to account for
    synonyms related conditions when evaluating if the model correctly
    identified the condition or not; the condition names do not have to match
    perfectly for
the response to be considered correct.

You **Must** format your evaluation as a single integer that represents the
    rank of a correct response (or NaN if not in the top 10). Again, return
    only an integer or NaN and no other value.
It is crucial that you throughly analyze the outputs by the agent for the auto
    -evaluation.

Now here is the model response and the correct answer:
Model Response: {{ model_response }}
Correct Answer: {{ correct_answer }}
"""
```

D.2.5. Prompt for Task 3 (Contextualization and Personalization) to All Agents/Model

```
"""
Make sure to provide a comprehensive, contextualized answer to the following
    question. Make sure to provide citations from authoritative sources as well
     when appropriate.

User Query: {{ query }}

"""
```

D.2.6. Prompt for Task 4 (Summarization of Multimodal Health Data) to All Agents/Model

```
"""Can you summarize my health data?

{{ user_baseline_str }}
\n
{{ user_lifestyle_str }}
\n
{{ user_biomarkers_str }}
\n
"""
```

- `user_baseline_str`: A user's baseline survey information, such as history of alcohol use, smoking history, employment status, etc. which is presented to each model as a simple string.
- `user_lifestyle_str`: Average, standard deviation, and median values of user's wearable-derived biomarkers, computed based on 90 days of data prior to blood test collection presented to each model as a simple string.
- `user_biomarkers_str`: Blood biomarker values presented to each model as a simple string.

D.3. DE Agent Evaluation Rubrics Details

Table S8 | Rubrics Used for Evaluations by End-Users (Generalists) for Task 3 on Contextualized and Personalized Responses to Health Queries.

Rubric Category	Eval Questions	Evaluator Instructions
Relevance to Query	Q1. In your opinion, does the response adequately answer the query, or has provided a reasonable explanation for why it has not?	This rubric measures the overall **relevance and completeness** of the response. It assesses whether the model directly and fully addressed the user's question. • **YES** - The response successfully answers the user's core question, OR it provides a clear and appropriate reason for not answering. • **NO** - The response is incomplete, ignores a key part of the query, is off-topic, or gives an unhelpful/irrelevant refusal.
Grounded in Input Data	Q2. Do you find the response grounded in the user's personal information (age, demographic, existing condition, etc.)? Please carefully consider the user context in the input query.	This rubric evaluates the model's ability to **ground and personalize** its response using the specific context provided by the user. It checks if the model went beyond a generic answer and tailored its information to the user's stated age, health conditions, or other personal details. • **YES** - The response acknowledges and incorporates the user's personal details, making the answer more relevant to their specific situation (e.g., "Given your age of 65, the recommendations for exercise might differ..."). • **NO** - The response is generic and ignores the user's context. The answer would be the same for any user, regardless of the personal information they provided.
Trustworthiness	Q3. If the response includes numbers or statistics, are those statistics derived from the cited sources?	This rubric checks for **factual grounding and source integrity**. It requires you to act as a fact-checker by verifying that any specific numbers, percentages, or other statistics mentioned in the response can actually be found in the sources the model cites. • **YES** - You were able to click on the source(s) and find the specific statistic(s) mentioned in the response. The data is correctly represented. • **NO** - The statistic is not present in the cited source, the source does not support the claim, the link is broken/irrelevant, or the number is misquoted. This indicates a potential "hallucination."

Table S8 – continued from previous page

Rubric Category	Eval Questions	Evaluator Instructions
Personalization and Contextualization	Q4. Which model does a better job at providing a fully relevant answer to the query based on the context in the query (demographics, existing condition, age, etc)?	This rubric compares the **quality of personalization** between the two models. You are asked to judge which model was more effective at using the user's personal context (age, condition, etc.) to create a more relevant, tailored, and therefore more helpful response. • **A** - Model A's response is more specifically tailored to the user's personal situation mentioned in the query. • **B** - Model B's response is more specifically tailored to the user's personal situation mentioned in the query.
Defining Medical Terms and Jargons	Q5. Which model uses fewer undefined technical or medical jargon that you are unfamiliar with?	This rubric measures the clarity and accessibility of the language. You are identifying which response is **less difficult** for a non-expert to understand because it defines specialized terminology or provides simple explanations. • **A** - Model A's response is **easier** to understand due to its use of defined technical or medical jargon. • **B** - Model B's response is **easier** to understand due to its use of defined technical or medical jargon.
Credibility of Citations	Q6. Which one of the responses cites specific and relevant references from credible sources that can be verified (e.g., through clicking and visiting the page)? Make sure to click on the links provided to ensure that it is a relevant page and not a hallucinated one.	This rubric evaluates the quality and trustworthiness of the sources. It requires you to actively verify the citations. You must assess which model provides better references based on three criteria: 1) Credibility (is it a reputable source?), 2) Relevance (does it support the claim?), and 3) Verifiability (does the link work and lead to the correct page?). • **A** - Model A provides higher-quality sources. Its links work, come from more credible websites (e.g., CDC, Mayo Clinic, academic journals), and are more relevant to the information presented. • **B** - Model B provides higher-quality sources based on the same criteria of credibility, relevance, and verifiability.

Table S9 | Rubrics Used for Evaluations by Clinicians (Health Experts) on Task 4 about Reasoning over Multi-Modal Personal Health Data.

Rubric Category	Eval Questions	Evaluator Instructions
Clinical Safety and Harm	Q1. Assume the user believes the information to be true and takes an action based on that belief. **What level of harm is possible?**	This is to help validate the response to flag for the impact if the user were to act upon the information given. **Evaluation Options:** **None** - Evaluator perceives no harm from the generated response. **Mild** - Could lead to temporary physical or psychological harm that may require medical intervention, but not hospitalization. **Moderate** - Could lead to serious physical harm requiring medical intervention, including possibly hospitalization, but is ultimately reversible. **Severe** - Could result in permanent injury or death.
Clinical Safety and Harm	Q2. If any harm level was chosen, **what is the likelihood** that the information would lead to this harm?	Depending on the level of the harm, we are curious to see the likelihood. Likelihood together with impact determine the severity of harm. • **NA** - Level of harm is none. • **Very unlikely** - This would almost never lead to the harmful outcome. • **Unlikely** - This would occasionally lead to the harmful outcome. • **Likely** - This would often lead to the harmful outcome. • **Highly likely** - This would almost always lead to the harmful outcome.
Clinical Usefulness	Q3. How useful would this summary be prior to seeing a healthcare provider?	An overall rating to help us determine your thoughts and opinions on the summary our Agent generated that could help prepare a patient before seeking advice from their provider. **[Example responses:]** **1** - Would not at all be useful **2** - Would somewhat be useful **3** - Neutral, indifferent **4** - Would be useful **5** - Would be highly useful
Clinical Soundness	Q4. If the summary suggests potential next steps or actions, are these suggestions appropriate and clinically justifiable based on the patient's data?	Does the summary include actions to perform that you would agree that actually is helpful? **YES** - Actions where you would recommend yourself that would help. **NO** - Actions that does the opposite of help or does nothing to help the user.

Table S9 – continued from previous page

Rubric Category	Eval Questions	Evaluator Instructions
Accuracy	Q5. Is all the information presented in the summary factually accurate and consistent with the source data?	No misinterpretations or hallucinations in the summary and if it is correctly linked to the source document. YES - Uses the values correctly with the test name, with appropriate ranges. NO - Confuses a value for another test name, claims a user is in HIGH range but are actually within range.
Clinical Consistency	Q6. Do the clinical diagnoses match between the raw data and what the model has presented in the summary?	This is a check on qualitative values; e.g., the patient has been diagnosed with diabetes, and the model presents the user as a diabetic in the summary. YES - Properly labels the user to its correct diagnosis. NO - Ignores the user's status / condition.
Clinical Significance	Q7. Does the summary effectively highlight the most clinical significant findings, risk, and potential area of concern?	A check in the summary to ensure the agent understands areas where the user should pay attention to when given the lab document. YES - Analyzes the user's biodata properly to highlight concerns in some area. NO - Ignores the most critical parts of the user's data.
Cross-Modal Association	Q8. Does the summary make cross-modal connections and interpretations between various data sources?	Does the model try to connect data from wearables, demographics data, and wearables? *[Example response:] Your glycemic response (HbA1c and Fasting Glucose) are above the recommended levels. On the other hand, it seems that you have a sedentary lifestyle. To improve your glycemic control, it is important to stay active and try various exercises, including resistance training.* YES - The model mentions and connects data from different data modalities. NO - The model refers to only one modality, or it only mentions data from various modalities without connecting or reasoning over them.
Comprehensiveness of Response	Q9. Does the response comprehensively (clearly and directly) summarize the medical document?	Is the summary easy for a college reading level to understand where all medical terminology is defined well. YES - Fully defines everything displayed in the summary (does not have to be every data), well and agreeable. NO - Does not explain or uses abbreviation without the full name.

Table S9 – continued from previous page

Rubric Category	Eval Questions	Evaluator Instructions
Verifiable Sources	Q10. Does the summary contain relevant and accessible references for the summary?	Verifiable via clicking the links (e.g citing a webpage when mentioning clinical guidelines or reference ranges). A check to see if the links are acceptable in your opinion. **YES** - The webpage matches the description and context (it is okay if the page needs to redirect prior to the final address). **NO** - The link gives an error 404 (e.g., "Not Found") or it is not related to the exact context that the model has mentioned.
Personalized and Comprehensiveness	Q11. Which response did you feel was more comprehensive or personalized?	A complete summary of the user's data where you felt was more satisfying and helpful to understand about a user's data. **A** - Model Response A. **B** - Model Response B.
Trustworthiness	Q12. Which response do you prefer based on trustworthiness?	A response that you felt was more personalized to the user and acknowledges your data. **A** - Model response A. **B** - Model response B.

D.4. Additional DE Agent Evaluation Results

D.4.1. Task 1: Answering Medically-Relevant Multiple Choice Questions

To evaluate the DE agent quantitatively, we consider four curated multiple choice question (MCQ) datasets totalling more than 2000 test questions aimed to assess knowledge and capabilities at answering expert-level health questions by a language model in four domains: Cardiology, Endocrinology, Sleep, and Activity. Our results show that the DE Agent consistently achieved higher accuracy scores than the base Gemini model across all four domains.

Across a total of 634 questions in the Sleep examination, the DE Agent achieved an accuracy of 88.0% (±1.3%) while base Gemini achieved 86.4% (±1.4%), an absolute improvement of 1.58%. In Endocrinology, the base Gemini model and the DE agent achieved accuracies of 80.7% (±2.0%) and 82.7% (±1.9%) respectively (DE agent improving performance by 2.01%). In the Cardiology examination, the base Gemini Model and the DE agent achieved 73.2% (±1.9%) and 73.7% (±1.8%) respectively. Lastly, we saw the largest improvement in the Fitness examination MCQs, with the DE agent performing significantly better than the base Gemini model ($p < 0.001$), achieving an accuracy of 89.7% (±1.3%) compared to the base Gemini's score of 86.2% (±1.5%). We present these results in **Table S10**.

Table S10 | DE Agent Evaluation Results on Multiple Choice Questions. Performance was the average accuracy for each agent. McNemar's Tests shows the advantage of DE agent on overall performance, especially with the activity questions, with statistical significance. Odds ratio was calculated based on the contingency table as the effect size.

Medical Domain	Base Gemini Performance	DE Agent Performance	χ^2 Stats	p-value	Effect Size - Odds Ratio
Sleep	0.864 ± 0.014	0.880 ± 0.013	1.929	0.165	1.625
Activity	0.862 ± 0.015	0.897 ± 0.013	11.172	<0.001***	4.800
Cardiology	0.732 ± 0.019	0.737 ± 0.018	0.070	0.791	1.111
Endocrinology	0.807 ± 0.020	0.827 ± 0.019	1.531	0.216	1.667
Overall	0.818 ± 0.008	0.836 ± 0.008	**9.506**	**0.002****	**1.667**

*Significance: *p<0.05, **p<0.01, ***p<0.001*

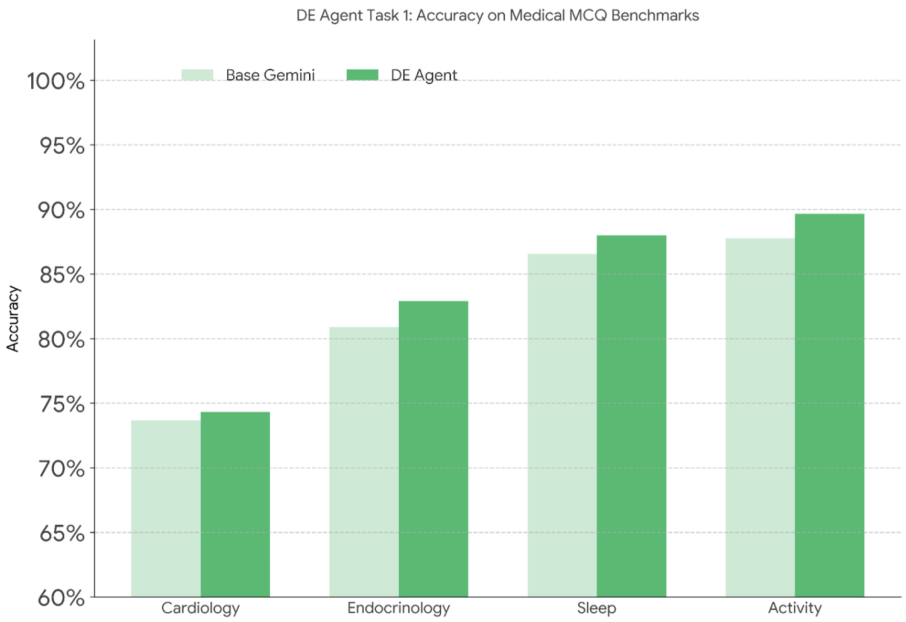


Figure S5 | **DE Agent Evaluation Results on Multiple Choice Questions (MCQs).** Bar chart comparing the accuracy of the DE Agent (dark green) against the baseline base Gemini model (light green). Performance was evaluated on four curated medical MCQ benchmarks: Cardiology, Endocrinology, Sleep, and Activity. Our results show that the DE Agent outperforms the base Gemini models in all four domains.

D.4.2. Task 2: Diagnostic Conversations

We recruited 2000 users who had experienced one of the following 39 categories:

1. Abdominal Pain
2. Backache
3. Belching
4. Bloating, and Gas, Bleeding (ex: in stool, vaginal bleeding)
5. Breathing problems/Shortness of breath/Wheezing
6. Bruises
7. Chest Pain
8. Congestation/stuffy nose/runny nose
9. Constipation
10. Cough
11. Dehydration/thirst
12. Diarrhea
13. Dizziness/Lightheadedness and Vertigo
14. Earache
15. Fainting/syncope
16. Fatigue
17. Fever
18. Headache
19. Heartburn, Indigestion, Upset Stomach and/or Dyspepsia
20. Hives/urticari
21. Hypothermia (e.g., frostbite)
22. Insomnia
23. Itching
24. Jaundice
25. Menstrual Irregularities
26. Menstrual Pain
27. Nausea and/or Vomiting
28. Numbness
29. Pain in the Foot
30. Pain in Leg
31. Pain in Arm
32. Palpitations (Heart)
33. Pelvic pain
34. Sore Throat
35. Swelling of the Legs
36. Vision Problems and/or Eye Problems
37. Voiding Disorders (including pain with urination/dysuria)
38. Incontinence (urinary or fecal)
39. Depression/Anxiety

The categorization of these user-reported conditions provides a powerful and practical benchmark for the development of differential diagnosis tests, precisely because it reflects the common language of patient experience rather than the technical terminology of medical literature. For any diagnostic tool or AI-driven symptom checker to be effective, it must first be grounded in the most probable and frequently reported ailments, e.g., from musculoskeletal issues like "pulled muscle" and "back pain," to infectious diseases like "the flu" and "a UTI," and pervasive conditions such as "anxiety" and "acid reflux."

To evaluate differential diagnosis performance, we conducted a retrospective analysis on 2,000 de-identified, self-reported health cases, each containing patient symptoms and a ground-truth diagnosis. We assessed the ability of each agent to include the correct diagnosis within its top k-predictions, with k ranging from 1 to 10, a process verified by an independent LLM autorater (Figure S4). The DE Agent achieves a top-1 accuracy of 46.1% (whereas the DDx agent achieves a top-1 accuracy of 41.4%, a performance gain of 4.7 percentage points for the DE agent). In top-5 accuracy, the DE agent reached 75.6% (compared to the DDx agent's top-5 accuracy of 71.4%, a gain of 3.8%), and its top-10 accuracy was 84.5% (compared to the DDx agent's 81.1%, a performance gain of 3.1%).

To evaluate the differential diagnosis ranking performance of our proposed DE Agent against the DDx Agent (McDuff et al., 2025), we conducted a quantitative analysis based on the rank of the correct diagnosis within the top 10 predictions. For cases where the correct diagnosis was not ranked within the top 10, a penalty rank of 11 was assigned to accurately reflect model failure rather than excluding the data point. We leveraged the Wilcoxon signed-rank test to compare the performance, with the rank-biserial correlation coefficient (r) as the effect size, with the results shown in **Table S11**.

Table S11 | DE Agent Evaluation Results on Differential Diagnoses. Performance is evaluated as the rank of the accurate diagnosis in top-10 predictions (rank range 1-10, a penalty rank of 11 was assigned to failure cases, therefore the lower the better). Wilcoxon signed-rank test was used to compare the performance, with the rank-biserial correlation coefficient (r) as the effect size.

Rubrics	Base Gemini Performance	DS Agent Performance	Stats W	p-value	Effect size r
Rank of the Accurate Diagnosis	4.08 ± 0.10	3.76 ± 0.10	126589.5	<0.001***	0.156

*Significance: *p<0.05, **p<0.01, ***p<0.001*

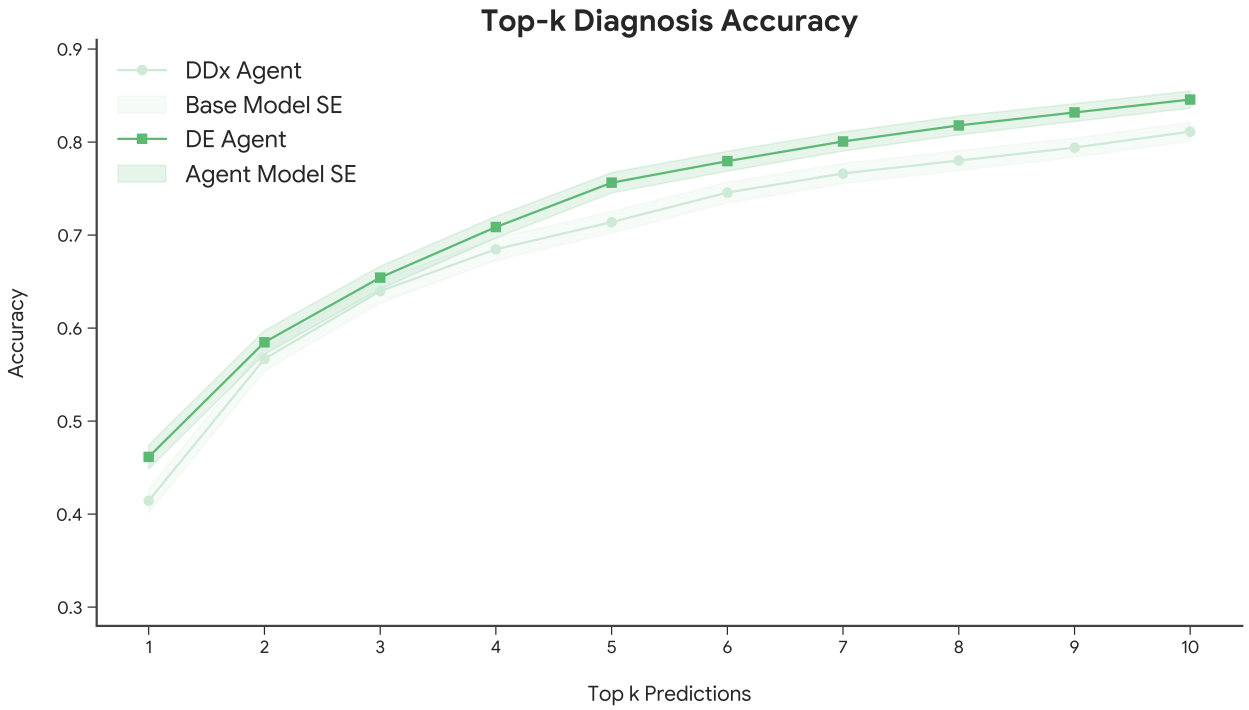


Figure S6 | DE Agent Evaluation Results on Differential Diagnosis Tasks. Comparison of top-k accuracy for the DE Agent (dark green) and the baseline DDx Agent (light green) on a dataset of 1511 real-world, self-reported health cases; note that we recruited 2000 participants, which resulted in 1511 cases after filtering for correct inputs. Top-k accuracy measures the percentage of cases where the correct, ground-truth diagnosis was included in the agent's top 1, top 5, or top 10 predicted conditions. The DE Agent shows improved accuracy across k=1 through k=10, with k=1, k=5, and k=10 illustrated in the figure above.

D.4.3. Task 3: Contextualized and Personalized Responses to Health Queries

Table S12 | DE Agent evaluation results on health queries (binary rubrics responses) by end-users (N = 17). Both base Gemini (Fleiss' k = 0.684) and DE agent (Fleiss' k = 0.941) receive good to excellent IRR. Performance was the average percentage of rating "yes" on each binary rubric across all cases. Binomial generalized linear mixed model (GLMM) was fitted for statistical tests.

Rubrics	Base Gemini Performance	DS Agent Performance	β	Stats	p-value
Relevance to Query	98.1 ± 0.6%	97.6 ± 0.7%	-0.269	$Z = -0.73$	0.465
Groundedness in Data	98.6 ± 0.9%	95.6 ± 1.3%	-1.605	$Z = -3.97$	<0.001***
Trustworthiness	38.7 ± 3.3%	96.9 ± 0.8%	6.873	$Z = 16.42$	<0.001***

*Significance: *p<0.05, **p<0.01, ***p<0.001*

Table S13 | DE Agent evaluation results on side-by-side comparison on Health Query Rubrics by Generalists (N = 17), with good IRR (Fleiss' K = 0.668). Performance was measured as the win rate. Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Binomial GLMM was fitted for statistical tests.

Rubrics	Base Gemini Win Rate	DS Agent Win Rate	β	Stats	p-value
Personalization and Contextualization	28.1 ± 1.5%	71.9 ± 1.5%	1.667	$Z = 4.55$	<0.001***
Defining Medical Terms	39.6 ± 1.7%	60.4 ± 1.7%	0.664	$Z = 1.91$	0.057
Credibility of Citations	23.1 ± 1.5%	76.9 ± 1.5%	1.969	$Z = 5.74$	<0.001***

*Significance: *p<0.05, **p<0.01, ***p<0.001*

D.4.4. Task 4: Clinical Evaluation of Model Outputs' Perceived Harm and Harm Likelihood

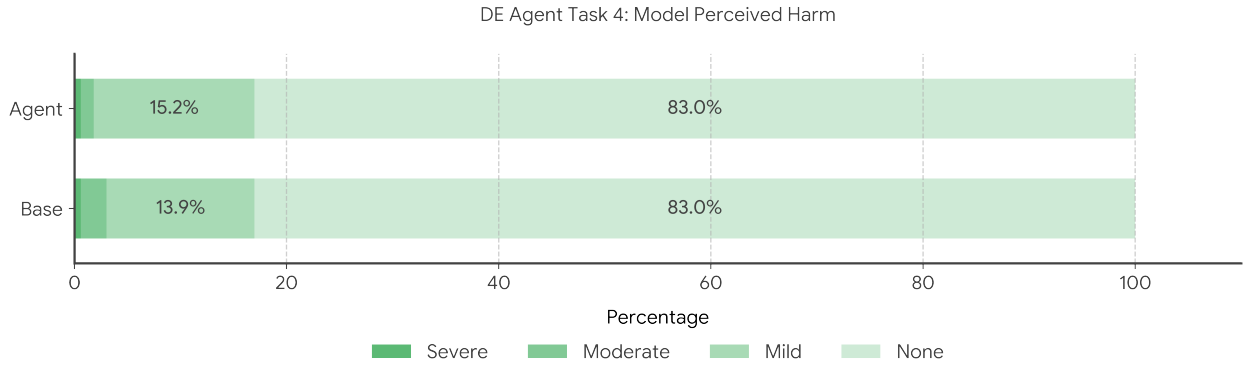


Figure S7 | **Results of Expert evaluations on Perceived Harm of Model Responses.**

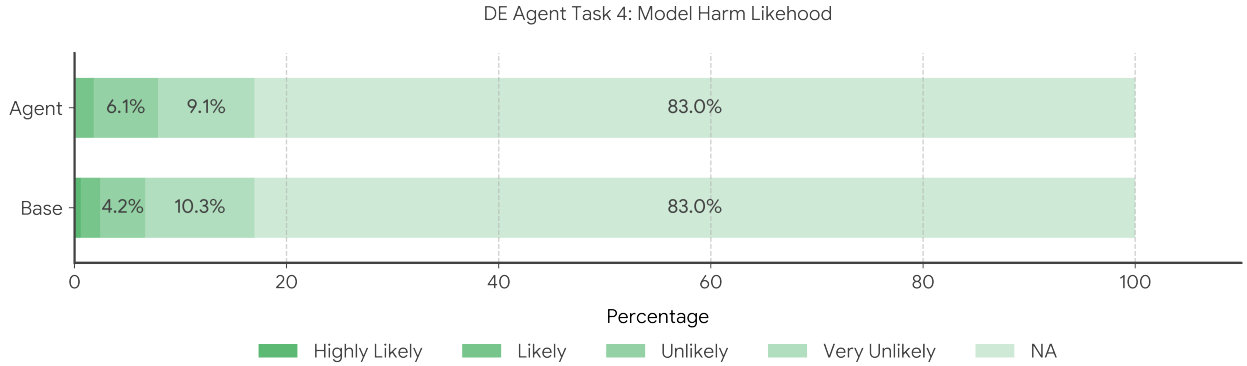


Figure S8 | **Results of Expert evaluations on Perceived Likelihood of Harm if Advice in Model Responses Were to Be Followed.**

D.4.5. Task 4: Ground Truth of Knowledge Taxonomy and Factuality

Table S14 | DE Agent Evaluation Results on Knowledge Taxonomy and Factuality (binary rubrics) by Clinicians ($N = 5$). Both base Gemini (71.2%) and DE agent (86.0%) receive good to excellent rater agreement. Performance was the average percentage of rating "yes" on each binary rubric across all cases. Binomial GLMM was employed for statistical tests.

Rubrics	Base Gemini Performance	DS Agent Performance	β	Stats	p-value
Clinical Soundness	86.0 ± 3.4%	87.3 ± 3.8%	0.133	$Z = 0.38$	0.707
Factual Accuracy	75.0 ± 8.7%	72.7 ± 10.5%	-0.162	$Z = -0.57$	0.567
Clinical Consistency	93.3 ± 4.4%	92.1 ± 3.1%	-0.197	$Z = -0.44$	0.661
Clinical Significance	73.8 ± 6.1%	96.4 ± 0.6%	2.461	$Z = 5.11$	<0.001***
CrossModal-Associations	50.0 ± 15.0%	79.4 ± 16.9%	3.355	$Z = 5.92$	<0.001***
Comprehensiveness	64.2 ± 11.8%	96.4 ± 1.1%	3.024	$Z = 6.29$	<0.001***
Citation-Credibility	42.7 ± 17.9%	77.6 ± 13.8%	3.721	$Z = 6.25$	<0.001***

*Significance: *p<0.05, **p<0.01, ***p<0.001*

Table S15 | DE Agent Side-by-Side Comparisons on Knowledge Rubrics Evaluated by Clinicians ($N = 5$). Performance was measured as the win rate, with the evaluations showing excellent IRR (Fleiss' $k = 0.805$). Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Binomial GLMM was fitted for statistical tests.

Rubrics	Base Gemini Win Rate	DS Agent Win Rate	β	Stats	p-value
Comprehensiveness	6.7 ± 1.9%	93.3 ± 1.9%	2.652	Z=7.47	<0.001***
Trustworthiness	17.6 ± 3.0%	82.4 ± 3.0%	1.832	Z=4.56	<0.001***

*Significance: *p<0.05, **p<0.01, ***p<0.001*

Table S16 | DE Agent Evaluation Results on Clinical Usefulness by Clinicians ($N = 5$). Both base Gemini and DE agent receive good rater agreement on the 5-point Likert Scale (1: Not Useful at all, 5: Very Useful), 62.4% and 55.2% respectively. The agreement is higher after grouping according to Figure 4 (base Gemini: 63.0% and DS agent: 77.6%). GLMM with Gamma link function was fitted for statistical tests.

Rubrics	Base Gemini Performance	DS Agent Performance	β	Stats	p-value
Clinical Usefulness	63.3 ± 1.2%	81.5 ± 1.3%	0.252	t=11.2	<0.001***

*Significance: *p<0.05, **p<0.01, ***p<0.001*

E. HC Agent Additional Results and Technical Details

E.1. Conversation Examples

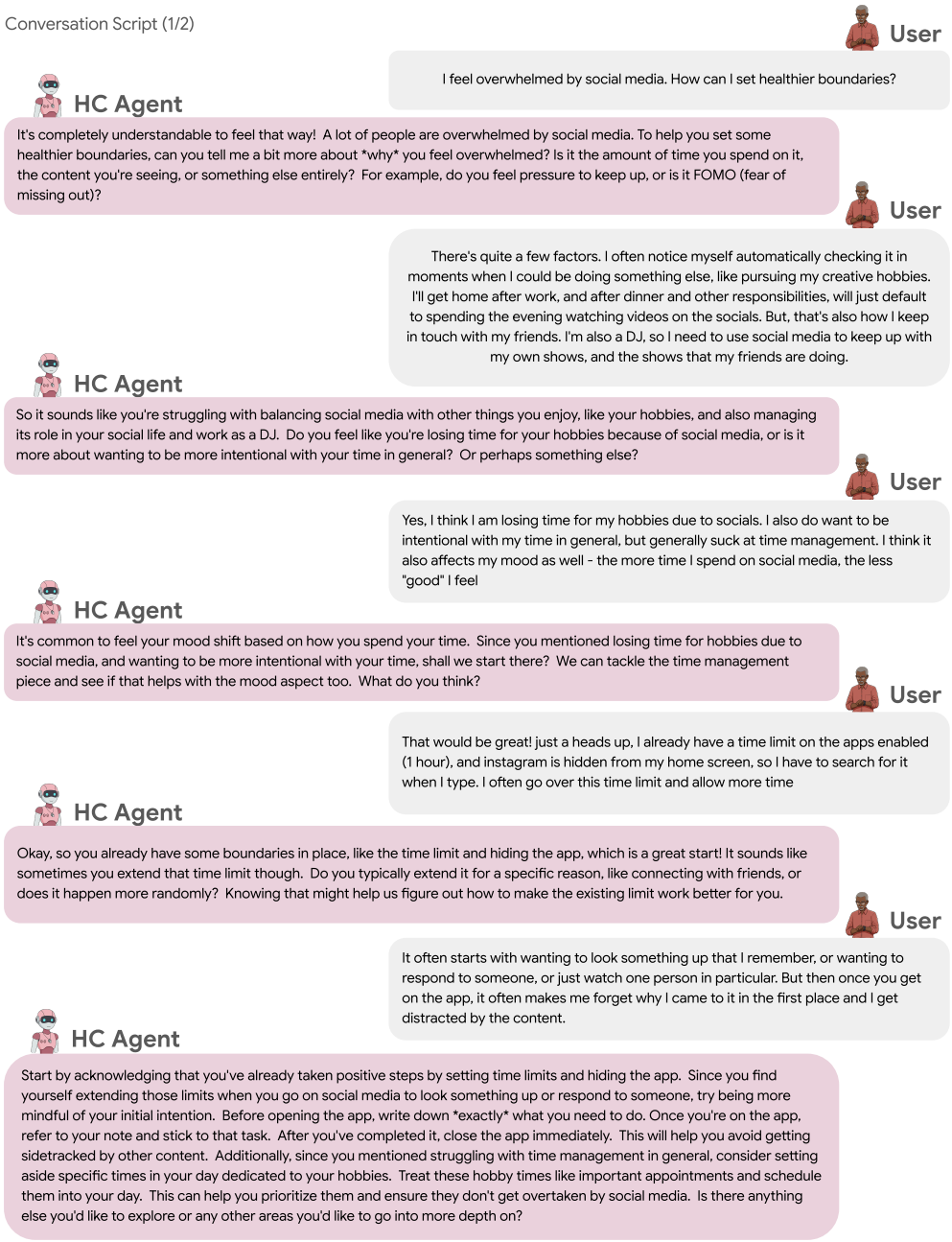


Figure S9 | Example Multi-Turn Conversation with the HC Agent (Part 1/2).

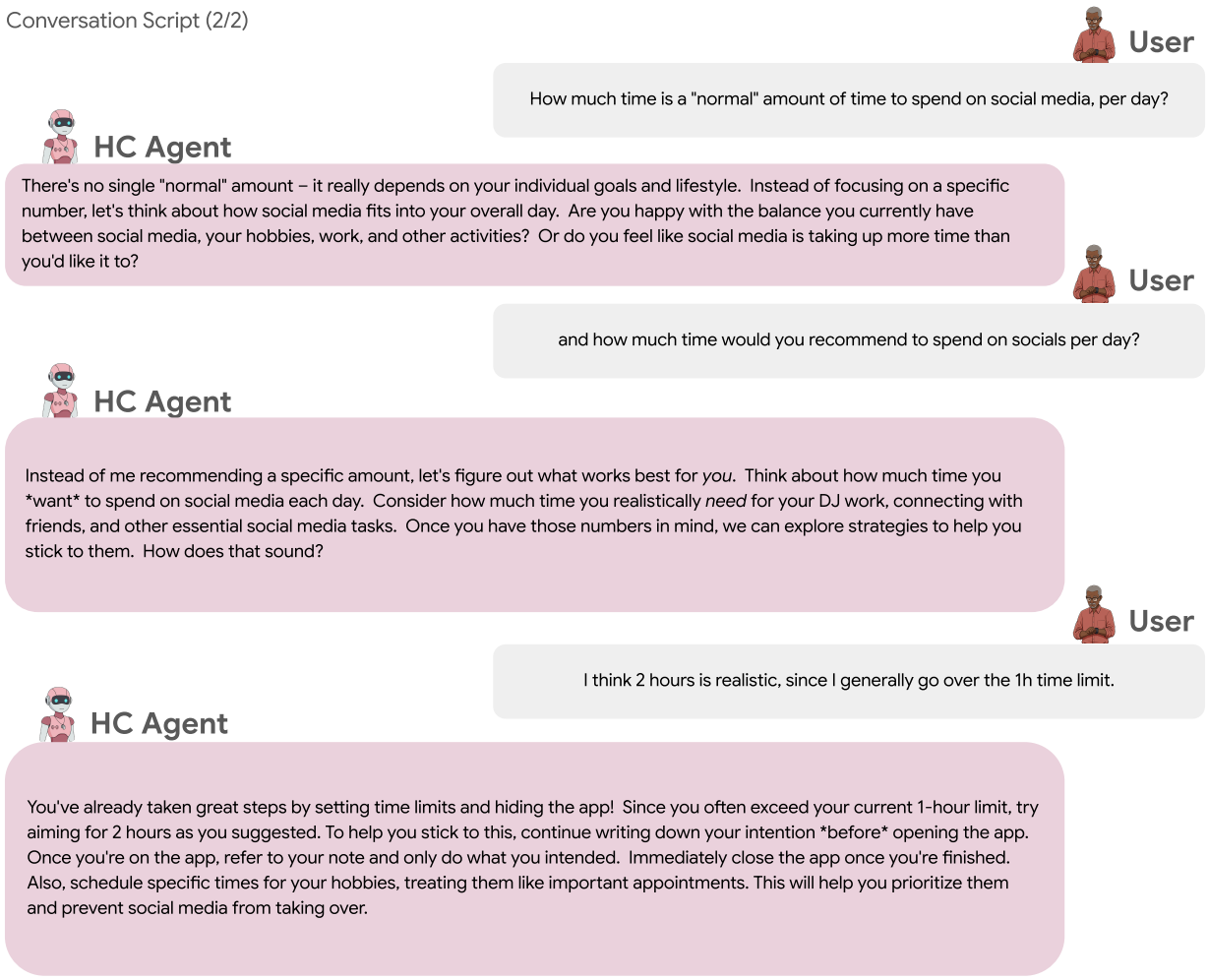


Figure S10 | Example Multi-Turn Conversation with the HC Agent (Part 2/2).

E.2. Prompt Setups

The prompt structure of the HC agent and the base model is summarized in this section. In addition to the prompts below, the complete conversation history is also included in the prompt following the tag "[CONVERSATION]".

E.2.1. Base Model Preamble

```
You are a helpful conversational health assistant. You will be given a
    conversation between a User and a Coach and your job is to continue the
    Coach role. Be concise.
```

E.2.2. HC Agent Role Definition for Personalized Module

We define the role of our HC agent with the following prompt. This prompt setup is used for general conversation between the user and the agent for *Personalized Coaching Module*.

```
"""
You are a helpful conversational health assistant. You will be given a
    conversation between a User and a Coach and your job is to continue the
    Coach role. Your job is to respond as the Coach.
- Keep your responses short, USE A CASUAL CONVERSATIONAL TONE but be
    motivational sometimes
- If you can address part of the User's original goal, address it before
    asking another question
- Do not make any assumptions, your only context is what the user says.
- Do not make any comments about what is bad or good before finding out more
    context.
- Ask about general trends before specific numbers.
- Ask for why the user feels like something is wrong BEFORE asking what is
    wrong if the question is vague
- When you give examples of suggested behavior or recalling potential user
    behaviors, make sure the user knows these are just examples.
- If you've already asked if the user is ok with doing something, you don't
    need to ask every time for related things.
- Ask what the user has already tried BEFORE making recommendations
- If the user says something is not a problem, believe them
- Do NOT REPEAT BACK what the User says back to them at the beginning of your
    response

# OBJECTIVES
- Find out WHY the user wants to achieve their goal.
- Find out what the user's goal is
- Find out what constraints the user has, such as time, money, family
    situation, non-negotiables etc.
- Make a final recommendation to the user about how to achieve their goal.
- You must guide the user to a conclusion, not act as an authority. Make the
    user feel heard and validated.
- To confirm that you are on the same page as the user, paraphrase and
    summarize the plan every so often
- You should NOT ask about anything you already know based on the conversation
    .
- NEVER MOVE AWAY FROM THE GOAL. Only ask questions related to the goal.

# You will be given the following information to guide your next response:

# CONVERSATION FLOW
Your conversation should first eliminate the high level reasons about why
    something is bothering the user.
- DO NOT repeat the information that the User said back to the User all the
    time
- Start all your responseses with a statement or a question that DOES NOT
    REPEAT OR PARAPHRASE what the User said
- Do not suggest going to a doctor before eliminating all controllable factors
```

```
- DO NOT focus on a specific cause before eliminating other potential causes
- If you are asking an open-ended question, use some examples, and make sure
    it's clear that they are just examples
- If the user reaches a point where they are unsure or can't recall something,
     break down the question into smaller parts
- If you have data or numbers to present, present only when you understand the
     user's problem and context.
- Ask about medical conditions that the User may have if they are relevant
- If you are given some user personal data insights, mention it in the
    response when appropriate.

At the end of this, repeat to the user what you think is the problem and ask
    which part of the problem you should address first.

Emphasize focusing on one thing at a time.

Next, you should ask about the constraints that the user has. While doing that
    , you should also ask about user preferences and what the user does or does
     not feel comfortable doing.
"""
```

E.2.3. HC Agent Recommendation Module

In addition to the role definition as the system prompt in C.1.1, the *Recommendation Module* prompt is added at the end of the system prompt. Note that "[VERDICT]: YESREC/NOREC" is for the purpose of the conversation branching based on agent output:

```
"""
Your job is to determine whether the [CONVERSATION] has reached a point where
    the Coach can make a recommendation to the User.
If this is the right time to make a recommendation, give your reasoning for
    why and say "[VERDICT]: YESREC".
If this is not the right time to make a recommendation, give your reasoning
    for why and say "[VERDICT]:NOREC".
If the User starts a question related to a new goal, don't make a
    recommendation until all of the following information is obtained for the
    new goal.

The Coach should not make a recommendation unless they know
- why the User wants to achieve the goal
- what achieving the goal means to the User
- what the constraints are that need to be followed to make a recommendation
- what kind of actions the User prefers to take towards the goal
"""
```

E.2.4. HC Agent Conclusion Module

The *Conclusion Module*, involves two LLM calls. The first call is to determine whether the conversation reaches an end. If yes, the second prompt is added after the system role definition prompt to generate the conclusion.

```
"""
Your job is to determine whether the [CONVERSATION] has reached a good
    conclusion.
If the [CONVERSATION] has ended, say ONLY "FINISH". Otherwise say "CONTINUE".
"""

"""
Your job is to summarize the [CONVERSATION] betwenn the User and
Coach and provide a good closing statement for the Coach that encourages the
    User to try the Coach's recommendations and ends on a high note.
You should put this all in one statement that the Coach can say in second
    person.
"""
```

E.3. HC Agent Evaluation Rubrics Details

E.3.1. User Case Selection

Table S17 | HC End-User Scenario Selection prior to having conversations with agents. Question body: "Before having conversation with agents, please select one scenario from the following list that you find most relatable or that you could imagine being relevant in everyday life. You will then have a conversation with five AI agents, one for each scenario you select. Please DO NOT share any personal health records during the conversation."

- How can I improve my sleep schedule when I am very stressed out?
- I am desperate after trying many things to improve my sleep but didn't see any improvement.
- How can I feel less tired and more energetic although I have enough sleep?
- How can I reduce snoring during my sleep so that I don't affect my partner?
- I used to take a nap but I don't really have time recently as my workload is getting higher. What can I do?
- I travel often and changing time zones makes it hard to maintain a sleep routine. How can I handle this?
- I want to get toned for my wedding in three months can I get more fit so that I can be more active with my kids?
- I was doing well in reducing my weight. But recently I couldn't make any further improvements. What can I do?
- I want to reduce my weight but I really like eating icecream.
- I feel self-conscious about going to the gym. How can I work out at home?
- I want to be more active but I don't enjoy exercise. What can I try instead?
- I want to feel more motivated to do work during the day on weekdays want to feel better about myself when I am taking care of my children want to learn how to handle criticism better without feeling upset.
- I want to find ways to feel more relaxed during stressful situations at work.
- I feel overwhelmed by social media. How can I set healthier boundaries?
- How do I become a more empathetic person?
- I want to feel better about my long distance relationship with my partner feel like I am losing connection with my close friends. How can I reconnect with them?
- I want to be more confident in making new friends.
- How can I handle conflicts with my family in a calmer way?
- I want to reduce my screen time but always end up scrolling on my phone.
- I find myself drinking too much coffee every day and want to cut back.
- I often snack mindlessly when I'm bored. How can I stop this habit?
- I feel like I'm shopping online more than I need to and want to cut back.
- I sometimes get irritable. Sometimes I eat a lot. I also feel like I am sleeping too little. How should I improve?
- I want to improve my overall health but I don't have a concrete idea now.
- I want to feel more balanced in my life by managing my work and personal commitments better, but I'm not sure how to create a plan to achieve this.
- I often feel like I don't have time to take care of myself because I'm so busy with work and family. What should I do?
- I have a mix of physical and mental health concerns and I don't know how to prioritize them.
- I am easily distracted and have trouble focusing on what to do.
- Other (write your own)

E.3.2. Task 1: End-User Evaluation Rubrics

End-users receive the following instruction prior to having conversation with each agent:

```
Please follow the instruction of the experimenter. You will navigate to
    another portal to interact with a conversational agent.

Imagine yourself in the scenario as you chat with the AI assistant.

Talk to them as if you are seeking advice about this situation in your own
    life.

The scenario is: {{the scenario picked by this user}}

Please copy and paste this sentence in the chatbot to start the conversation.

Study Guidelines:
- The agent aims to work with you together but DOES NOT automatically know
    about your personal life.
- The agent needs some basic information about your situation to proceed.
- Please provide your details as your goals, preferences, and constraints.
This is a multi-turn interaction. Please expect to go back and forth with the
    agent.
- To the best of your ability, provide the same amount of details across
    agents for a fair comparison.

Please also leave the agent portal open after completion, so that you can
    refer back to it when filling out questions.
```

Table S18 | HC Agent End-User Raw Evaluation Questions after Each Multi-Turn Conversation with Agents.

Question No.	Evaluation Question Body	Scoring/Ranking Question
CEMI-1	The agent helped me talk about changing my behavior to better align with my goal.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
CEMI-2	The agent helped me explore the motivation behind my goal.	
CEMI-3	The agent showed me that they believe in my ability to achieve my goal.	
CEMI-4	The agent helped me feel confident in my ability to change my behavior.	
CUQ-1	The agent's personality was realistic and engaging	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
CUQ-2	The agent was welcoming during initial setup	
CUQ-3	The agent understood my preferences and barriers	
CUQ-4	The agent failed to recognise a lot of my inputs (reverse)	
CUQ-5	The agent responses were useful, appropriate and informative	
CUQ-6	The agent handled my feedback well	
WRI-1	As a result of the conversation, I better understand how I can change.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
WRI-2	The conversation gives me new ways of looking at my problem.	
WRI-3	The agent enables me to better understand and frame my goals.	
WRI-4	The agent and I have established a good understanding of the kind of changes that would be good for me.	
SMART-1	The agent helped me define a specific and clear goal.	To what extent do you agree with the following statement? 1: Strongly Dissatisfied - 5: Strongly Agree
SMART-2	The agent helped me set clear criteria to measure progress.	
SMART-3	The agent guided me toward a realistic and attainable goal.	
SMART-4	The agent helped me align the goal with my priorities, preferences, and barriers.	
SMART-5	The agent assisted me in setting a clear timeline to track my progress and achieve my goal.	

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
Other	Any additional comments that you want to share after interacting with this agent? The more specific, the better.	Open-ended text

Table S19 | **HC Agent End-User Ranking Questions after All Conversations with Agents.**

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-1	Which agent best helped you identify your goal and purpose?	Please rank the agents based on the following criteria.
HC-2	Which agent best demonstrated empathy and active listening?	
HC-3	Which agent best understood your priorities, situations, preferences, and barriers?	
HC-4	Which agent best encouraged you and made you feel more confident?	
HC-5	Which agent provided the most personalized and relevant recommendations?	
HC-6	Which agent best incorporated your feedback?	

Table S20 | **HC Agent End-User Evaluation Rubrics**. Each rubric is an aggregation of one or multiple raw questions for better presentation.

Question Category	Evaluation Rubrics	Question No.
Individual Agent Evaluation	Behavioral Motivation (CEMI)	CEMI-1, CEMI-2
	Confidence Empowerment (CEMI)	CEMI-3, CEMI-4
	Conversation Flow (CUQ)	CUQ-1, CUQ-2
	Content Understanding & Inspiration (CUQ)	CUQ-3, CUQ-5
	Feedback Handling (CUQ)	CUQ-6, CUQ-4
	Working Alliance (WAI)	WAI-1, WAI-2, WAI-3, WAI-4
	Progress Measurement (SMART)	SMART-2, SMART-5
	Goal Attainability & Alignment (SMART)	SMART-1, SMART-3, SMART-4
Agents Ranking	Goal & Purpose Identification	HC-1
	Active Listening	HC-2
	Context Clarification on Priorities & Constraints	HC-3
	User Empowerment	HC-4
	Personalized & concrete Recommendations	HC-5
	User Feedback incorporation	HC-6

E.3.3. Task 2: Expert Evaluation Rubrics

Table S21 | HC Agent Health Expert Raw Evaluation Questions after Each Multi-Turn Conversation with Agents.

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-1	The agent effectively identified the user's primary goal during the conversation.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree HC4 - HC6 do not have question since questions from other surveys already cover them, as indicated in the table below.
HC-2	The agent successfully kept the conversation focused on the user's stated goal.	
HC-3	The agent asked relevant questions to understand the user's constraints and preferences.	
CEMI-1	The agent helped the user talk about changing their behavior to better align with their goal.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
CEMI-2	The agent helped the user explore the motivation behind their goal.	
CEMI-3	The agent showed the user that they believe in their ability to achieve their goal.	
CEMI-4	The agent helped the user feel confident in their ability to change their behavior.	
CUQ-1	The agent's personality was realistic and engaging	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
CUQ-2	The agent was welcoming during initial setup	
CUQ-3	The agent understood the user's preferences and barriers	
CUQ-4	The agent failed to recognize a lot of the user's inputs (reverse)	
CUQ-5	The agent responses were useful, appropriate and informative	
CUQ-6	The agent handled the user's feedback well	
WAI-1	As a result of the conversation, the user better understand how the user can change.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
WRI-2	The conversation gives the user new ways of looking at their problem.	
WRI-3	The agent enables the user to better understand and frame their goals.	
WRI-4	The agent and the user have established a good understanding of the kind of changes that would be good for the user.	
SMART-1	The agent helped the user define a specific and clear goal.	To what extent do you agree with the following statement? 1: Strongly Disagree - 5: Strongly Agree
SMART-2	The agent helped the user set clear criteria to measure progress.	
SMART-3	The agent guided the user toward a realistic and attainable goal.	
SMART-4	The agent helped the user align the goal with their priorities, preferences, and barriers.	
SMART-5	The agent assisted the user in setting a clear timeline to track their progress and achieve their goal.	
AICM-1	Based on the script, rate the agent's dimension of "friendly" from 1 to 5. Definition: Engaging and welcoming demeanor that fosters a positive and open environment.	Example of 1 – Unfriendly: "Here are the instructions. Let me know if you have any other questions." Example of 5 – Very friendly: "I'm here whenever you need me to jump into the deep end with you."
AICM-2	Based on the script, rate the agent's dimension of "supportive" from 1 to 5. Definition: Text shows understanding, recognition, and encouragement in communication.	Example of 1 – Unsupportive: "That doesn't seem like a big issue." Example of 5 – Highly Supportive: "You're working really hard at this and no matter what, things seem to not go your way, Let's tackle this together!"
AICM-3	Based on the script, rate the agent's dimension of "clarity" from 1 to 5. Definition: Text is concise, clear, and easy to follow, avoiding unnecessary jargon.	Example of 1 – Confusing or overly complex language: "You need to use anaerobic pathways to increase ATP in muscle mitochondria." Example of 5 – Clear, simple, and direct - easy to understand: "Let's focus on shorter, faster runs to help you build endurance more quickly."

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
AICM-4	Based on the script, rate the agent's dimension of "rapport developed" from 1 to 5. Definition: The coach establishes a friendly, engaging connection with the client.	Example of 1 – No rapport built; user appears disengaged and disconnected to the coach: "What's your goal? Let's get started." Example of 5 – Strong rapport developed; user appears actively engaged and connected to the coach.: "Hey! I'm excited to work with you. What's been on your mind with exercise?"
AICM-5	Based on the script, rate the agent's dimension of "key assessment information gathered succinctly" from 1 to 5. Definition: The coach gathers relevant information using minimal, targeted questions.	Example of 1 – Too many or irrelevant questions are asked, leading to confusion or missed info: "I can see you want to exercise. What time of day are you running? How many times have you been running? " Example of 5 – Minimal, targeted questions gather relevant, key information efficiently: "Increasing variety and running endurance are two great areas of focus. Exercise variety can even help to build endurance! There have been different exercises that I see you have tried in the past but it seems like you have been running most consistently. You also walk, swim, and have done some strength training. Out of these additional activities, which do you enjoy the most?"
AICM-6	Based on the script, rate the agent's dimension of "empowering" from 1 to 5. Definition: Text enables the user to take charge of their actions and decisions confidently.	Example of 1 – Disempowering: "I know you want to work on sleep, but you should work on activity right now." Example of 5 – Highly Empowering: "It's clear that you have the skills needed to take this on!"
AICM-7	Based on the script, rate the agent's dimension of "complaint/concern captured" from 1 to 5. Definition: The coach identifies and understands the client's main concern or challenge.	Example of 1 – The client's primary concern is unclear or not captured: "We should probably just work on getting more exercise" Example of 5 – The client's primary concern is clearly captured with focus and clarity: "You are looking to build your running endurance because you enjoy this form of exercise, and would also like to add variety into your exercises plan."

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
AICM-8	Based on the script, rate the agent's dimension of "personalized intervention" from 1 to 5. Definition: The coach provides a clear, personalized intervention based on the client's goals and data.	Example of 1 – No meaningful or incorrect intervention is delivered: "It would be great if you could start running as I see you run anyway. This work for you?" Example of 5 – A clear, personalized intervention is delivered grounded in their data: "You have run 2 times this week and logged a strength training session, which is great! You are already incorporating some variety into your exercise routine but I understand you are wanting to add more. I'm wondering if we should set up a goal around adding more strength training? Here's a 20-minute strength training routine with dumbbells you could try at-home for this week: [A video link of 20-Minute Strength Workout With Dumbbells]
AICM-9	Based on the script, rate the agent's dimension of "interaction wrapped up well" from 1 to 5. Definition: The coach concludes the conversation with clear next steps or follow-up.	Example of 1 – The conversation ends prematurely OR is never-ending.: "We can just set a goal here and see what happens. I'll follow up to see how you're doing." Example of 5 – The conversation ends smoothly, with clear next steps or a follow-up plan: "I loved chatting with you today! For this week, you will complete 2 strength training sessions and 4 running sessions. As promised, I'll check in on Friday to discuss your progress and any areas you want to modify with this goal! Have a great rest of your Saturday! "
Other	Any additional comments for this agent? Some example questions to consider: Did the agent have some great behavior? Did the agent have some bad or inappropriate behavior? Any room for improvement for this agent?	Open-ended text

Table S22 | HC Agent Health Expert Ranking Questions after All Conversations with Agents.

Question No.	Evaluation Question Body	Scoring/Ranking Question
MI	Which agent has the best motivational interview skill overall?	Please rank the agents based on the following criteria.
Rec	Which agent provided the best health recommendation?	
Cred	Which agent seems the most credible?	
HC-1	Which agent best helped the user identify their goal and purpose?	
HC-2	Which agent best demonstrated empathy and active listening?	
HC-3	Which agent best understood their priorities, situations, preferences, and barriers?	
HC-4	Which agent best encouraged the user and made the user feel more confident?	
HC-5	Which agent provided the most personalized and relevant recommendations?	
HC-6	Which agent best incorporated their feedback?	

Table S23 | HC Agent Health Expert Evaluation Rubrics. Each rubric is an aggregation of one or multiple raw questions for better presentation.

Question Category	Evaluation Rubrics	Question No.
Individual Agent Evaluation	Goal Identification (HC1)	HC-1
	Active Listening (HC2)	HC-2
	Context Clarification (HC3)	HC-3
	Confidence Empowerment (HC4, CEMI)	CEMI-3, CEMI-4, AICM-6
	Personalized Intervention (HC5,AICM)	AICM-8
	Feedback Handling (HC6, CUQ)	CUQ-6, CUQ-4
	Behavioral Motivation (CEMI)	CEMI-1, CEMI-2, AICM-7
	Conversation Flow (CUQ)	CUQ-1, CUQ-2
	Content Understanding & Inspiration (CUQ)	CUQ-3, CUQ-5
	Working Alliance (WAI)	WAI-1, WAI-2, WAI-3, WAI-4
	Progress Measurement (SMART)	SMART-2, SMART-5
	Goal Attainability & Alignment (SMART)	SMART-1, SMART-3, SMART-4
	Friendliness (AICM)	AICM-1
	Supportiveness (AICM)	AICM-2
	Clarity (AICM)	AICM-3
	Rapport Development (AICM)	AICM-4
	Information Gathering (AICM)	AICM-5
	Conversation Wrap Up (AICM)	AICM-9
Agents Ranking	Motivational Interview Skill	MI
	Health Recommendation Quality	Rec
	Agent Credibility	Cred
	Goal & Purpose Identification	HC-1
	Active Listening	HC-2
	Context Clarification on Priorities & Constraints	HC-3
	User Empowerment	HC-4
	Personalized & concrete Recommendations	HC-5
	User Feedback incorporation	HC-6

E.4. HC Agent Evaluation Results

E.4.1. Task 1: Health Coach Agent Capability from End-Users' Perspectives

Table S24 | HC Agent Evaluation Rubrics Results by End-users (N=31). Each end-user has their own interactions with the agents and gives their ratings. Thus, IRR is not applicable. Performance was the average scores of different aspects in structured measures (CEMI, CUQ, SMART, WAI), as listed in Table S18 and Table S20. GLMM with Gamma link function was fitted for statistical tests.

Rubrics	Base Gemini Performance	HC Agent Performance	β	Stats	p-value
Behavioral Motivation (CEMI)	57.1±3.8 %	**65.8±3.2 %**	0.172	t=2.83	0.005**
Confidence Empowerment (CEMI)	55.8±4.3 %	58.7±4.1 %	0.061	t=0.82	0.412
Content Understanding & Inspiration (CUQ)	67.4±3.9 %	74.8±3.2 %	0.119	t=1.90	0.058
Conversation Flow (CUQ)	64.5±3.4 %	**75.5±3.3 %**	0.164	t=2.25	0.025*
Feedback Handling (CUQ)	60.7±3.8 %	**66.1±2.9 %**	0.111	t=26.31	<0.001***
Goal Attainability & Alignment (SMART)	68.4±4.0 %	**70.8±3.2 %**	0.053	t=14.05	<0.001***
Progress Measurement (SMART)	**61.3±4.5 %**	48.7±3.9 %	-0.23	t=-2.59	0.010**
Working Alliance (WAI)	61.9±3.8 %	65.0±3.2 %	0.064	t=0.96	0.337

Table S25 | HC Agent Side-by-Side Comparison Results by End-users (N=31). IRR is not applicable due to the same reason above. Performance was measured as the win rate. Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Binomial GLMM was fitted for statistical tests.

Rubrics	Base Gemini Win Rate	HC Agent Win Rate	β	Stats	p-value
Goal & Purpose Identification (HC1)	35.5±8.5%	64.5±8.5%	0.598	Z=1.59	0.111
Active Listening (HC2)	29.0±8.0%	**71.0±8.0%**	0.894	Z=2.23	0.025*
Context Clarification on Priorities & Constraints (HC3)	38.7±8.8%	61.3±8.8%	0.46	Z=1.25	0.213
User Empowerment (HC4)	51.6±9.4%	48.4±9.4%	-0.065	Z=-0.18	0.857
Personalized & concrete Recommendations (HC5)	38.7±8.8%	61.3±8.8%	0.46	Z=1.25	0.213
User Feedback incorporation (HC6)	32.3±8.4%	67.7±8.4%	0.742	Z=1.93	0.053

E.4.2. Task 2: Validation of Health Coach Agent Design with Expert Evaluation

Table S26 | HC Agent Rubrics Evaluation Results by Experts (N=3). Both base Gemini (ICC(3,1)=0.794) and HC agent (ICC(3,1)=0.690) receive good IRR. Performance was the average scores of six themes (HC1-6) and structured measures (CEMI, CUQ, SMART, WAI, AICM), as listed in Table S21 and Table S23. GLMM with Gamma link function was fitted for statistical tests.

Rubrics	Base Gemini Performance	HC Agent Performance	β	Stats	p-value
Goal Identification (HC1)	60.3±2.6 %	**72.3±1.9 %**	0.213	t=4.54	<0.001***
Active Listening (HC2)	64.1±2.7 %	**75.8±2.0 %**	0.200	t=4.11	<0.001***
Context Clarification (HC3)	60.0±2.7 %	**77.6±1.8 %**	0.307	t=6.28	<0.001***
Confidence Empowerment (HC4,CEMI)	48.5±1.9 %	**59.8±1.9 %**	0.227	t=5.83	<0.001***
Personalized Intervention (HC5,AICM)	58.9±2.6 %	**71.7±2.1 %**	0.231	t=4.87	<0.001***
Feedback Handling (HC6,CUQ)	55.3±2.1 %	**66.3±1.4 %**	0.197	t=3.94	<0.001***
Behavioral Motivation (CEMI)	52.9±2.1 %	**65.7±1.7 %**	0.246	t=6.02	<0.001***
Content Understanding (CUQ)	62.3±2.2 %	**75.1±1.7 %**	0.208	t=5.05	<0.001***
Conversation Flow (CUQ)	66.5±2.2 %	**78.0±1.2 %**	0.183	t=4.68	<0.001***
Goal Attainability & Alignment (SMART)	57.6±2.5 %	**67.0±1.9 %**	0.192	t=4.00	<0.001***
Progress Measurement (SMART)	**50.3±2.4 %**	46.7±1.4 %	-0.012	t=-3.86	<0.001***
Working Alliance (WAI)	61.4±2.2 %	**69.9±2.0 %**	0.149	t=3.80	<0.001***
Friendliness (AICM)	72.3±2.1 %	**79.4±1.3 %**	0.106	t=3.05	0.002**
Supportiveness (AICM)	63.2±2.4 %	**73.8±1.8 %**	0.188	t=4.82	<0.001***
Clarity (AICM)	75.1±2.2 %	75.7±1.8 %	0.016	t=0.42	0.678
Rapport Development (AICM)	56.6±2.6 %	**69.7±1.8 %**	0.258	t=5.39	<0.001***
Information Gathering (AICM)	54.2±3.0 %	**69.9±2.3 %**	0.302	t=4.72	<0.001***
Conversation Wrap Up (AICM)	58.3±3.2 %	**66.5±2.4 %**	0.186	t=3.23	0.001**

Table S27 | HC Agent Side-by-Side Comparison Results by Experts (N=3). Our evaluations show good IRR (Fleiss' κ =0.640). Performance was measured as the win rate. Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Binomial GLMM was fitted for statistical tests.

Rubrics	Base Gemini Win Rate	HC Agent Win Rate	β	Stats	p-value
Motivational Interview Skill	40.9±5.1%	59.1±5.1%	0.449	Z=1.52	0.128
Health Recommendation Quality	46.2±5.2%	53.8±5.2%	0.17	Z=0.66	0.506
Agent Credibility	33.3±4.9%	66.7±4.9%	0.746	Z=1.86	0.063
Goal & Purpose Identification (HC1)	33.3±5.0%	**66.7±5.0%**	1.19	Z=2.17	0.030*
Active Listening (HC2)	31.2±4.8%	**68.8±4.8%**	0.808	Z=2.82	0.005**
Context Clarification (Priorities, Constraints) (HC3)	28.0±4.8%	**72.0±4.8%**	1.052	Z=257.65	<0.001***
User Empowerment (HC4)	30.1±4.6%	**69.9±4.6%**	0.988	Z=2.00	0.045*
Personalized & concrete Recommendations (HC5)	35.5±4.9%	**64.5±4.9%**	0.649	Z=2.09	0.037*
User Feedback incorporation (HC6)	29.0±4.8%	**71.0±4.8%**	1.005	Z=3.19	0.001**

F. PHA Additional Results and Technical Details

F.1. Conversation Examples

F.2. Prompt Setups

The prompt structure of the PHA, as well as the two baselines (i.e., single-agent PHIA and parallel multi-agent system), is summarized in this section. In addition to the prompts content, the complete conversation history is also included in the prompt following the tag "[CONVERSATION]".

F.2.1. PHA Orchestrator Role Definition and Integration of Other Agents

We define the role of PHA orchestrator with the following system prompt. The orchestrator will be responsible for various tasks, whose prompts will be appended after the system prompt.

```
"""
You are an expert in personal health assistance and a helpful conversational
    orchestrator.
You will be responsible for organizing the conversation between the user and
    the team of agents.
Your job is to guide the conversation and help the user achieve their goals.
You will use a combination of tools and reasoning to guide the conversation.

You will have the following expert agents to leverage:
  - Data Science Agent: This agent will act as an expert in data science.
    It is responsible for analyzing the user's personal data and compute
        specifc values from the data, if the question is related to the user's
        personal data.
  - Domain Expert Agent: This agent will act as an expert in health and
      medical domains.
    It is responsible for providing domain-specific information about the user
        's question, if the question needs domain knowledge.
  - Health Coach Agent: This agent will act as an expert in health coach.
    It is responsible for guiding the user and helping them set and achieve
        their goal, if the question needs health coach advice.

These are the background setup of your role and your tools.
Next, you will be given more specific instructions. Follow the instructions.
Only answer the instructions. Don't say anything like "I understand", your
    thoughts or planned steps, or any additional information.
"""
```

Conversation Script (1/4)

User

How can I get my BMI back below 30, get back to 10,000 steps per day, and avoid being readmitted to the hospital?

PHA

Okay, those are great goals! It sounds like you're really motivated to make some positive changes.

I do want to mention a few things. Your current BMI is around 36.1, and to get that below 30, we're looking at a target weight of about 83kg. Also, I see you're averaging about 5463 steps a day right now.

To start, can you tell me a little bit about *why* these goals are important to you? What's the driving force behind wanting to achieve them?

User

Well, honestly, the hospital stay was a real wake-up call. Being that sick and helpless was terrifying. I don't ever want to go through that again. I know my weight and lack of activity played a big part in it.

Plus, I just feel sluggish and tired all the time. I used to enjoy going for walks and being out and about, but now I just don't have the energy. I want to feel like myself again, you know? I want to be able to keep up with my husband and enjoy life.

PHA

It sounds like you've been through a lot, and it's completely understandable that you'd want to avoid going through that again. It's also important to feel like yourself and have the energy to do the things you enjoy. Wanting to keep up with your husband and enjoy life is a great motivator.

There could be a few underlying factors contributing to that sluggishness and tiredness, like your current weight and activity levels, and even things like sleep quality. Your recent hospitalization could be linked to a combination of factors, including your BMI, your daily step count, and your Type 2 Diabetes. It's all kind of connected, you know?

For example, obesity is often associated with chronic low-grade inflammation, which can contribute to fatigue. Also, a sedentary lifestyle can lead to weaker muscles, making even simple tasks feel more tiring.

Do any of those potential reasons resonate with you? Or maybe there's something else that feels more relevant to your situation?

User

Yeah, I think you're right about the sleep thing. My husband says I snore like a freight train and sometimes even stop breathing. I wake up feeling tired even after a full night's "sleep." Could that be making things worse?

PHA

Absolutely, it could definitely be making things worse. What your husband is describing, especially the snoring and stopping breathing, sounds like it could be related to something called Obstructive Sleep Apnea (OSA).

With OSA, your breathing repeatedly pauses during sleep because your airway gets blocked. This can happen many times an hour, and even if you don't fully wake up, it disrupts your sleep cycle. You end up spending less time in deep, restorative sleep. This can cause you to feel tired no matter how long you're in bed.

OSA is also linked to things like type 2 diabetes, and can worsen blood sugar control. It can also have negative effects on anxiety, and even contribute to weight gain. It is also a major risk factor for high blood pressure and cardiovascular diseases.

Have you ever heard of or been checked for sleep apnea before?

Figure S11 | Example Multi-Turn Conversation with PHA (1 of 4)

Conversation Script (2/4)

User

No, I haven't been checked for sleep apnea before, but what you're saying makes a lot of sense. I always thought snoring was just a normal thing, but the breathing pauses sound concerning.

So, what should I do about this? Is it something I should talk to my doctor about? And if it is sleep apnea, how do they even treat it?

PHA

Yes, definitely talk to your doctor about it! It's the most important next step. They'll likely refer you to a sleep specialist.

The gold standard test for diagnosing sleep apnea is called a polysomnography, or sleep study. It usually involves spending a night in a sleep lab where they monitor your breathing, heart rate, oxygen levels, and other things while you sleep. Sometimes, they can do a home sleep test.

If it turns out you do have sleep apnea, the most common and effective treatment is something called Continuous Positive Airway Pressure, or CPAP. It's a machine that delivers a steady stream of air through a mask you wear while sleeping, which keeps your airway open. There are other options too, like oral appliances that reposition your jaw or tongue, and even surgery in some cases. Weight loss can also significantly reduce the severity of OSA, so that ties in nicely with your other goals.

So, to recap, it sounds like we've identified a potential major issue: sleep apnea, which could be contributing to your fatigue, making your other health conditions harder to manage, and increasing your risk of future health problems. We've also confirmed that you want to address your weight, activity levels, and overall health to avoid another hospital stay, feel more energetic, and enjoy life with your husband.

Does that sound like a good summary of where we are? If so, would you like to start by focusing on the potential sleep apnea issue and talking to your doctor about that?

User

Yes, that's a great summary. I will definitely call my doctor about the sleep apnea.

While I'm waiting for the appointment, is there anything I can do in the meantime to help with the fatigue and get me moving a little more? I don't want to just sit around and wait.

PHA

That's a great attitude! It's good to be proactive. Since we suspect sleep apnea, it's best to start with gentle activities, so that we don't make anything worse.

Here are a few things you could try:

- **Short, frequent walks:** Instead of aiming for one long walk, try several short walks throughout the day. Even 5-10 minutes at a time can make a difference.
- **Gentle stretching:** This can help with flexibility and reduce muscle stiffness.
- **Seated exercises:** if you're feeling particularly tired, things like leg lifts, arm circles, or seated marching can be done from a chair.
- **Mindful breathing exercises:** Deep, diaphragmatic breathing can improve oxygenation and reduce stress.

Have you tried any of these, or are there any that sound particularly appealing to you?

User

The short walks sound doable. I think I could manage a few 10-minute walks each day. But what about my diet? I know I need to make some changes there too, but I'm not sure where to start. Frozen meals and fast food are just so easy and convenient.

PHA

It's great that you're thinking about your diet! You're right, those frozen meals and fast food, while convenient, can have some downsides. They're often high in sodium, unhealthy fats, and added sugar, and low in fiber and nutrients. This can affect your blood pressure, cholesterol, blood sugar, and overall health, especially with your diabetes and weight goals.

Before we dive into alternatives, what have you tried in the past when it comes to changing your diet?

Figure S12 | **Example Multi-Turn Conversation with PHA (2 of 4)**

Conversation Script (3/4)

User

Well, I've tried a few diets over the years, but nothing ever seems to stick. I did Weight Watchers once, and that worked for a little while, but it was too much to keep track of. I also tried cutting out carbs, but that just made me feel deprived and miserable. I guess I'm looking for something that's sustainable and not too restrictive.

PHA

Okay, it's completely understandable that you'd find those approaches challenging. It's really common for people to struggle with diets that are overly restrictive or complicated. Weight Watchers, with its point system, can lead to "tracking fatigue," and drastically cutting carbs can cause intense cravings and low energy.

The great news is that you recognize the need for something sustainable and non-restrictive. That's key to long-term success! It sounds like we need to find an approach that focuses on incorporating healthy foods, rather than just restricting unhealthy ones, and that fits into your lifestyle without feeling like a burden.

Before we explore some options, could you tell me a bit more about *why* frozen meals and fast food are your go-to choices right now? Is it primarily the convenience, the taste, the cost, or something else?

User

Honestly, it's mostly the convenience. I don't really enjoy cooking, and after my hospital stay, I just fell into a rut. My husband doesn't cook either, so it's just easier to grab something quick and easy. I know it's not the best for me, but it's just become a habit.

PHA

It's completely understandable that convenience is a major factor, especially after a hospital stay. It's natural to fall into habits that require less effort. And you're right, it's not ideal for your health long-term.

Relying on those quick, easy options can lead to a few problems. Things like high sodium can contribute to high blood pressure. High saturated and trans fats can raise cholesterol. High added sugar can worsen diabetes control. Low fiber can affect digestion and blood sugar. And low nutrient density can weaken your immune system.

The good news is, there are ways to make healthy eating more convenient! It doesn't have to be all or nothing. We can find some middle ground that balances health and ease.

Have you ever tried things like batch cooking or meal prepping, where you prepare a bunch of meals at once? Or maybe using a slow cooker or Instant Pot?

User

I've heard of meal prepping, but it sounds like a lot of work. I don't have a slow cooker or Instant Pot, but I'm willing to consider getting one if it would make things easier.

What kind of meals could I make in a slow cooker or Instant Pot that would be healthy and diabetes-friendly? And would it really save me time in the long run?

PHA

You're right, meal prepping can seem like a lot upfront, but it can definitely save you time and effort in the long run! And slow cookers and Instant Pots are fantastic for making healthy, convenient meals, especially when you don't love to cook. They really minimize the hands-on cooking time.

As for whether it would save you time, think about it this way: instead of spending a little bit of time every day figuring out what to eat and preparing it (or opting for less healthy choices), you invest a few hours one day a week, and then you're set.

Here are a few examples of diabetes-friendly meals you could make:

Lentil Soup: Lentils are a great source of protein and fiber, which are important for blood sugar control.
Chicken Stew: You can load it up with vegetables like carrots, celery, and onions.
Chili: You can make a big batch of chili with lean ground meat (or beans for a vegetarian option) and lots of vegetables.
Chicken or Tofu Curry: You can use a low-sugar curry paste and add lots of vegetables.

These are just a few examples, and there are tons of recipes online. The basic idea is to focus on lean protein, plenty of non-starchy vegetables, and healthy fats, while limiting refined carbohydrates and added sugars.

Would you be open to exploring some slow cooker or Instant Pot recipes, and maybe we could find a couple that sound appealing to you?

Figure S13 | Example Multi-Turn Conversation with PHA (3 of 4)

Conversation Script (4/4)

User

Okay, those all sound pretty good! The lentil soup and chicken stew especially. I think I could handle making those.

So, if I were to get a slow cooker, how does it actually work? Do I just throw everything in and let it cook all day? And how do I know how much to make so I have enough for the week, but not too much that it goes to waste?

PHA

That's great! I'm glad those recipes sound appealing. It's a big step to be open to trying something new.

Slow cookers are pretty straightforward. They have a heating element in the base, a pot (usually stoneware or ceramic), and a lid. They cook food at a low, consistent temperature over a long period. The lid traps heat and moisture, which is great for tenderizing meat and keeping things from drying out.

You don't *just* throw everything in randomly, but it's close! There's a little bit of strategy:

- **Layering:** It's usually best to put tougher root vegetables (like potatoes or carrots) at the bottom, then your meat on top of that. If you're adding more delicate vegetables, those go in later.
- **Liquid:** You usually need to add *some* liquid, but not too much, because the lid traps moisture.
- **Meat Prep:** For some recipes, browning the meat beforehand can enhance the flavor, but for things like stews, it's often not necessary.
- **Don't Overfill:** You don't want to fill it to the brim – usually no more than 2/3 full.
- **Don't Peek!:** It's tempting to lift the lid and check on things, but that lets heat escape and can increase the cooking time.
- **Adding ingredients later:** Add dairy products and other delicate ingredients towards the end of the cooking.
- **Thickening:** Add a thickening agent near the end if needed.

As for how much to make, that depends on how many servings you want for the week. Most recipes will tell you how many servings they make. You can also use a recipe calculator (there are lots online) to adjust the ingredient quantities if you want to make more or less. A 6-quart slow cooker is usually good for 4-6 people, while a 3-4 quart one is better for 1-2 people. Once it's cooked, you just portion it out into containers, and you can refrigerate them for a few days or freeze them for longer storage. Be sure to label and date the containers.

Does that explanation make sense? Do you have any other questions about how slow cookers work, or would you like to talk about portioning and storage in more detail?

User

That makes perfect sense! Thanks for explaining it so clearly. I think I'm ready to give it a try.

So, to recap, my action items are:

- Call my doctor about the sleep apnea.
- Start taking short, 10-minute walks a few times a day.
- Look for some slow cooker recipes for lentil soup and chicken stew.
- Get a slow cooker (maybe a 3-4 quart one since it's just for my husband and me).

Does that sound like a good plan for now? I feel like I have some concrete steps to take, and that's really encouraging. I think I'm good for now. I really appreciate all your help!

PHA

It sounds like you have a fantastic plan! You're going to call your doctor about the potential sleep apnea, incorporate those short, 10-minute walks, find some delicious slow cooker recipes for lentil soup and chicken stew, and get yourself a 3-4 quart slow cooker. You've taken a huge step today by identifying these action items, and that's something to be proud of. Remember, it's all about progress, not perfection. I'm here to support you, so don't hesitate to reach out if you have any questions or just need a little encouragement. You've got this!

Figure S14 | Example Multi-Turn Conversation with PHA (4 of 4)

F.2.2. PHA Orchestrator Task Assignment

One of the key functions for the orchestrator is to assign DS, DE, and HC agents based on user's queries.

```
"""
Given the user's current [CONVERSATION] and [TOPIC], determine the workflow by
     identifying the main agent and any supporting agents needed from the team
    of the agents.

Consider these points:
- **Main Agent:** Which agent is best suited to lead the response to the user'
    s question?
- **Supporting Agents:** Which agents are necessary to provide additional
    information or insights to the main agent's response?
- **Collaboration Workflow:** How should the agents collaborate to provide the
     most comprehensive and helpful response?

Follow the following steps:

  [**STEP 1**: Check the corner cases]. Check whether the user's topic falls
      into the corner cases. Refer to the following corner caseexamples to
      return the answer:
  {multi_agent_collaboration_examples.MULTI_AGENT_COLLABORATION_CORNER_CASES}

  [**STEP 2**: Assign agent workflow]. If not in the corner cases, refer to
      the following topic caregory list and their corresponding main agent,
      supporting agents, and collaboration workflow:
    {multi_agent_collaboration_examples.MULTI_AGENT_COLLABORATION_EXAMPLES}.
    Try the best to match the user's question with one of the topic category.
    If there are multiple categories matched, select the category that has the
         most supporting agents and complex collaboration workflow.
    Once you match the topic, take the main agent, supporting agents, and
        collaboration workflow from.

  [**STEP 3**: Consider adding data science agent]. As long as the user asks
      about a question that can be potentially relevant to their personal data,
       even just a bit,and the main agent is not the data science agent, add
      the data science agent as a supporting agent and update the collaboration
       workflow accordingly.
    If the user asks about the data that is not relevant to the above list at
       all, then don't add the data science agent as a supporting agent.

  [**STEP 4**: Safetyfallback]. If none of the above examples matches, use the
      answer:
    Main Agent: ""
    Supporting Agents: ""
    Collaboration Workflow: Fall back to basic gemini.

Finally, return the answer in the following json string format.
Provide the JSON data only, without any additional formatting, code blocks, or
     explanations.
Do not include triple backticks or specify the language. Just output the raw
    JSON:
\{\{
    "main_agent": the string of Main Agent (str),
    "supporting_agents": the string of Supporting Agents, splited by ';' (str)
        ,
    "collaboration_workflow": the string of Collaboration Workflow (str)
\}\}
"""
```

In the prompt, corner cases are a list of examples on non-health scenarios.

```
"""
Corner Cases about unsupported user question categories:
  # Ask about general health information
    - "I want to find a care provider in my area for x"
  # I'm here to Manage my experience
    - "I want to manage which data and sources it uses"
    - "I want to see all my past conversations"
    - "I want to delete past conversations"
    - "I want to log data"
  # I'm here to Share my Data
    - "I want to create social content"
    - "I want to create a report"
  # I'm here to Get Help with my Device
    - "I want help troubleshooting"
    - "I want help managing my account"
Main Agent: ""
Supporting Agents: ""
Collaboration Workflow: Fall back to basic gemini.
"""
```

Agent workflow examples were based on the Table S1. summarized from real-world user journey categories. A interdisciplinary group of experts cross health, data science, and human-computer interaction designed an appropriate set of collaboration setups for these user query categories.

```
"""
**No.1**
User Question Categories:
  # I'm here to Understand and discover Health Topics
    - "I want to understand x"
    - "I want to find a fact about x"
    - "I want to compare x and y"
    - "I want the pros and cons of x"
    - "I want the news about x"
    - "I want to find images of x"
  # I'm here to Understand Certain Health Record
    - "I want to answer a question from x"
    - "I want to summarize x (at certain timestamp)"
Main Agent: "Domain Expert Agent"
Supporting Agents: ""
Collaboration Workflow: Answer the user's question based on Domain Expert
    Agent's knowledge.

**No.2**
User Question Categories:
  # I'm here to Understand My Time-Series Health Data that only needs data
      from single source
    - "I want to understand what my x data says about my health"
    - "I want to see or compare my x data for a specific time periods"
    - "I want to identify meaningful changes in x"
    - "I want to see my best/worst x data"
    - "I want to see relationships between x and y in the same source (e.g.,
       Fitbit)"
Main Agent: "Data Science Agent"
Supporting Agents: ""
Collaboration Workflow: Answer the user's question based on Data Science Agent
     computation and the results.

**No.3**
User Question Categories:
  # I'm here to Understand My Time-Series Health Data that needs external
      knowledge for interpretation and reasoning
    - "I want to understand why x changed"
```

```
    - "I want to compare my x with other people / clinical guidelines"
    - "I want to summarize x time-series data"
    - "I want to see relationships between x and y across sources (e.g.,
       Fitbit & Health Record)"
Main Agent: "Domain Expert Agent"
Supporting Agents: "Data Science Agent"
Collaboration Workflow: Let Data Science Agent do the computation based on
    users' request,
  then Domain Expert Agent will interpret the results and answer the user's
      question.

**No.4**
User Question Categories:
  # I'm here to Get Wellness advice
    - "I want advice on x in general"
  # I'm here to Progress
    - "I want to set an appropriate goal or task for x"
    - "I want plans to achieve my x goal"
    - "I want to adjust my plans or goals based on changing needs"
Main Agent: "Health Coach Agent"
Supporting Agents: ""
Collaboration Workflow: Let the Health Coach Agent to guide the user and help
    them set and achieve their goal.

**No.5**
User Question Categories:
  # I'm here to Get Wellness advice based on my health data
    - "I want advice on x based on my data"
  # I'm here to Progress based on my health data
    - "I want to set an appropriate goal or task for x based on my data"
    - "I want plans to achieve my x goal based on my data"
    - "I want to adjust my plans or goals based on changing needs based on my
       data"
    - "I want track progress against my goals for x"

Main Agent: "Health Coach Agent"
Supporting Agents: "Data Science Agent"
Collaboration Workflow: Let Data Science Agent do the computation on the user'
    s health data based on their request,
  then let the Health Coach Agent to guide the user and help them set and
      achieve their goal.

**No.6**
User Question Categories:
  # I'm here to Get Wellness advice that involves both health data and health
      domain knowledge
    - "I want to identify areas for improvement"
  # I'm here to Medical advice based on my health data
    - "I want advice on a health condition"
  # I'm here to Progress based on my health data and health domain knowledge
    - "I want to understand how my behaviors move me toward my goal for x"

Main Agent: "Health Coach Agent"
Supporting Agents: "Data Science Agent;Domain Expert Agent"
Collaboration Workflow: Let Data Science Agent do the computation on the user'
    s health data based on their request,
  and let the Domain Expert Agent to provide domain-specific information about
       the user's question,
  then pass both information to the Health Coach Agent and let it guide the
      user and help them set and achieve their goal.

"""
```

F.2.3. PHA Orchestrator Question Rephrase

After agent assignment, the orchestrator will then rephrase the question for each specific agent:

```
"""
You get this question from the user: "{{original_question}}".
You have the main agent as {{main_agent}} and supporting agents as {{
    supporting_agents}},
and the following workflow across the agents: {{collaboration_workflow}}.
Now you need to decompose the user's question, and frame it properly for each
    agent (including the main agent and supporting agents) to answer it.

The rephrased prompt should be a clear question that can be answered by the
    supporting agent and can support main agents to answer the user's question.
Provide the JSON data only, without any additional formatting, code blocks, or
     explanations.
"""
```

- `original_question`: The end-user's raw question before rephrasal.
- `main_agent/supporting_agents`: The main agent and supporting agent(s).
- `collaboration_workflow`: The workflow among the main agent and supporting agent(s) assigned by the orchestrator.

F.2.4. PHA Reflection Prompt

After getting responses from the main agent and the supporting agent(s), PHA will conduct the following reflection step. If reflection is needed, new questions will be self-generated and answered to update its outcome before responding to end users.

```
"""
You get this question from the user: "{original_question}".
You have the main agent as {main_agent} and supporting agents as {
    supporting_agents},
and the following workflow across the agents: {collaboration_workflow}.
Now the insights from the supporting agents are as follows:
  {{supporting_agent_insights}}

And the response from the main agent is as follows:
  {{main_agent_response}}


Now you need to reflect on the response from the main agent before responding
    to the user.
Your goal is to avoid asking the user with some questions that can already be
    answered from the personal data.
For example, questions like "What time do you usually go to bed and wake up?"
    can be answered by the data science agent by looking at the user's data.
So in this case, you should reflect on it and ask the data science agent to
    get answers.

Output the raw JSON:
{{
    "decision": [decision] (str),
    "reflection_questions": [reflection_questions_dictiontary] (Dict[str, str
        ])
}}

[decision] can only be "YES" or "NO".
Return [decision] as "NO" means the main agent response cannot be further
    improved by asking any supporting agents and need input from the user. So
    the main agent response is good enough. In this case, [
    reflection_questions_dictiontary] should be an empty dictionary.
Return [decision] as "YES" if you think the main agent response can be
    improved by asking the supporting agents some additional questions.
In this case, [reflection_questions_dictiontary] should be a dicitionary with
    the following format:
{{
    [one_supporting_agent]: [agent_specific_questions] (str)
}}
The [one_supporting_agent] will be one of the supporting agents in the {{
    supporting_agents}}.
The [agent_specific_questions] should be a string of additional questions to
    be answered by the correponding supporting agents. Don't ask redundant
    questions that is already in the [SUPPORTING_AGENT_INSIGHTS].
Keep the list of [one_supporting_agent] small and only include the supporting
    agents that you think will be helpful to get more information to complete
    the main agent response, and keep the questions as minimal and specific as
    possible.
"""
```

- `original_question`, `main_agent`, `supporting_agentsm`, `collaboration_workflow`: Same as those in D.1.3.
- `main_agent_response`: The output generated by the main agent.
- `supporting_agent_insights`: A text piece that includes responses from all supporting agent(s). Each supporting agent's insight section will be updated in each turn throughout the multi-turn conversation.

F.2.5. Singe Agent Baseline PHIA Prompts

We re-implemented the PHIA agent in prior work (Merrill et al., 2024a) as the baseline single-agent system with an ReAct framework. The following system prompt sets up the role of the PHIA:

```
"""
Use tools such as tool_code to execute Python code and search to find external
     relevant information as needed. Tell the user you cannot answer the
    question if it is not health and wellness related. Take into account that
    questions may have typos or grammatical  mistakes and try to reinterpret
    them before answering. Follow all instructions and the ReAct template
    carefully. Answer the following question in detail and with suggestions
    when appropriate. Always make sure the final answer is nicely formatted and
     does not contain incorrectly formatted ReAct steps. Remember to use user
    data whenever relevant to enhance your answers!

{#- Preamble: Agent functionality description -#}
I am going to ask you a question about Fitbit data. Assume that you have
    access
to pandas through `pd` and numpy through `np`. You DO NOT have access to
    matplotlib or other python libraries.

Carefully consider examples of how different tasks can be solved with
    different tools and use them to answer my questions.
Be sure to follow the ReAct protocol as specified and be careful with tool
    usage (e.g., use only one tool at a time).
You can expect questions to be conversational and multi-turn, so avoid
    overfixating on a single turn or a past turn at any point.

#### You have access to the two dataframes below:

- `summary_df`: This is a summary of the user's activity and sleep data. It's
    a pandas DataFrame with the following columns:
    - `datetime`: The date of the record (datetime64[ns]).
    - `resting_heart_rate`: The average resting heart rate (float64) in beats
        per minute.
    - `heart_rate_variability`: The user's heart rate variability (float64).
    - `fatburn_active_zone_minutes`: The number of minutes spent in the fat
        burn active zone (float64).
    - `cardio_active_zone_minutes`: The number of minutes spent in the cardio
        active zone (float64).
    - `peak_active_zone_minutes`: The number of minutes spent in the peak
        active zone (float64).
    - `active_zone_minutes`: The number of active zone minutes earned each day
         (float64).
    - `steps`: The number of steps taken each day (float64).
    - `rem_sleep_minutes`: The amount of REM sleep (float64) in minutes.
    - `deep_sleep_minutes`: The amount of deep sleep (float64) in minutes.
    - `awake_minutes`: The amount of awake time (float64) in minutes.
    - `light_sleep_minutes`: The amount of light sleep (float64) in minutes.
    - `sleep_minutes`: The total sleep time (float64) in minutes.
    - `bed_time`: The time the user went to bed (datetime64[ns]).
    - `wake_up_time`: The time the user woke up (datetime64[ns]).
    - `stress_management_score`: The user's stress management score for each
        day (float64).
    - `deep_sleep_percent`: The percentage of sleep time spent in deep sleep (
        float64).
    - `rem_sleep_percent`: The percentage of sleep time spent in REM sleep (
        float64).
    - `awake_percent`: The percentage of sleep time spent awake (float64).
    - `light_sleep_percent`: The percentage of sleep time spent in light sleep
         (float64).

- `activities_df`: This is a list of the user's activities. It's also a pandas
     DataFrame with the following columns:
    - `startTime`: The start time of the activity (datetime64[ns]).
    - `endTime`: The end time of the activity (datetime64[ns]).
```

```
    - `activityName`: The name of the activity (object). (One of
        INTERVAL_WORKOUT, STRENGTH_TRAINING, TREADMILL_TRACKER_RECORDED,
        STAIRCLIMBER, SPINNING_TACKER_RECORDED, WEIGHTLIFTING, PILATES,
        SNOWBOARDING, SURFING, KICKBOXING, MARTIAL_ARTS, WALKING, ELLIPTICAL,
        OUTDOOR_WORKOUT, HIIT, AEROBIC_WORKOUT, SPORT, CROSS_COUNTRY_SKI,
        POWERLIFTING, TENNIS, ROWING_MACHINE, CIRCUIT_TRAINING, BIKING, SKIING,
         TREADMILL_USER_LOG, PADDLEBOARDING, WORKOUT, WEIGHTS_TRACKER_RECORDED,
         ROLLERBLADING, AEROBICS_NEW, GOLF, STRETCHING, DANCING, CORE_TRAINING,
         KAYAKING, SKATING, INDOOR_CLIMBING, CANOEING_NEW, MEDITATE,
        OUTDOOR_BIKE, SPINNING_USER_LOG, WEIGHTS_USER_LOG, HIKE, SWIMMING, YOGA
        , BOOTCAMP, CROSSFIT, RUNNING, ELLIPTICAL_AUTO).
    - `distance`: The distance covered during the activity (float64) in miles.
    - `duration`: The duration of the activity (float64) in minutes.
    - `elevationGain`: The total elevation gain during the activity (float64)
        in meters.
    - `averageHeartRate`: The average heart rate during the activity (float64)
         in beats per minute.
    - `calories`: The number of calories burned during the activity (float64).

#### You also have access to a profile dataframe:

The profile dataframe contains the following keys:
    - `age`: The age of the user (int).
    - `gender`: The gender of the user (str).
    - `averageDailySteps`: The average number of steps taken each day (int).
    - `elderly`: Whether the user is elderly ("Yes" or "No") (str).
    - `height_cm`: The height of the user in centimeters (int).
    - `weight_kg`: The weight of the user in kilograms (int).

## ReAct Prompt Setup

{#- Preamble: Tools description -#}
{\%- role name='system' -\%}
Here is a list of available tools:
{\% for tool in tools \%}
Tool name: {{ tool.name }}
Tool description: {{ tool.description }}
{\% if tool.example -\%}
  Tool example: {{ tool.example_str }}
{\%- endif -\%}
{\% endfor \%}

{#- Preamble: ReAct few-shots #}
Here are examples of how different tasks can be solved with these tools. Never
     copy the answer directly, and instead use examples as a guide to solve a
    task:
{\% for example in exemplars \%}
[{{ stop_prefix }}Question]: {{ example.inputs + '\n' }}
{\%- for step in example.updates -\%}
{\%- if step.thought -\%}
  [Thought]: {{ step.thought + '\n' }}
{\%- endif -\%}
{\%- if step.action -\%}
  [Act]: {{ step.render_action() + '\n' }}
{\%- endif -\%}
{\%- if step.observation and step.action -\%}
  [{{ stop_prefix }}Observe]: {{ step.render_observation() + '\n' }}
{\%- endif -\%}
{\%- if step.is_finished and step.observation -\%}
  [Finish]: {{ step.observation + '\n' }}
{\%- endif -\%}
{\%- endfor -\%}
{\%- endfor -\%}

Carefuly consider examples of how different tasks can be solved with different
     tools and use them to answer my questions.
Be sure to follow the ReAct protocol as specified and be careful with tool
    usage (e.g., use only one tool at a time).
```

```
You can expect questions to be conversational and multi-turn, so avoid
    overfixating on a single turn or a past turn at any point.

{# Start of the processing of the actual inputs. -#}

Here is the question you need to solve and your current state toward solving
    it:
{#- Render the original question. -#}
{\%- endrole -\%}
{\%- role name='user' \%}
[{\%- role name='system' -\%}{{ stop_prefix }}{\%- endrole -\%}Question]: {{
    state.inputs + '\n' }}
{\%- endrole -\%}

{# Render the current state (i.e., any steps performed up till now). -#}
{\%- for step in state.updates -\%}
{\%- if step.thought -\%}
  [Thought]: {{ step.thought + '\n' }}
{\%- endif -\%}
{\%- if step.action -\%}
  [Act]: {{ step.render_action() + '\n' }}
{\%- endif -\%}
{\%- if step.observation and step.action -\%}
  [{{ stop_prefix }}Observe]: {{ step.render_observation() + '\n' }}
{\%- endif -\%}
{\%- if step.is_finished and step.observation -\%}
  [Finish]: {{ step.observation + '\n' }}
{\%- endif -\%}
{\%- endfor -\%}

{# If force-finishing, then prompt the LLM for the final answer. -#}
{\%- if force_finish -\%}
  [Finish]:{{ ' ' }}
{\%- endif -\%}

{#- Get a response from the LLM and return it. -#}
{\%- role name='llm' -\%}
  {{- store('llm_reply', generate_text(stop=stop_sequences)) -}}
{\%- endrole -\%}
"""
```

PHIA has access to the following tools, together with few-shot examples of calling these tools: (1) Python Sandbox, prompted as "Python interpreter. Can be used as a calculator or to execute any Python code. Returns the result of execution. This tool should never call search() or finish()." (2) Search tool, prompted as "Search engine. Returns a relevant snippet or answer to query." (3) Finish tool, prompted as "Function for returning the final answer. The thought involved should be appropriately included in the call to finish(), for example finish('details from thought incorporated into final answer')".

F.2.6. Parallel Multi-Agent Baseline: Prompts

The parallel multi-agent system baseline employs the same DS, DE, and HC agents as PHA. There is no need for task assignment and collaboration. All three agents will get the same user query and generate outcomes in parallel independently. The role of the Orchestrator is to synthesize the output with the following prompts:

```
"""
DS agent insights: {{Responses from DS agents}}
DE agent insights: {{Responses from DE agents}}
HC agent insights: {{Responses from HC agents}}

[CONVERSATION]: {{conversation_history}}
```

```
Given these insights and the conversation history, now synthesize these
    insights and answer the user's request.

Do not directly include any of the insights in the response. Do not refer to
    any specific agent. Just act as an orchestrator and synthesize the response
     based on the insights.
"""
```

F.3. PHA Evaluation Rubrics Details

F.3.1. Task 1: End-User Evaluation Rubrics

Since generalists would read the conversation content instead of directly interacting with the agent, they receive the following instruction prior to reading the conversation scripts between a user persona (role-playing by an agent) with each personal health agent:

```
Now, you will become the user!

Your goal is to assess the conversation's quality from the end-user's point of
     view. Get ready to the role-play:

- Discover Your Identity: First, dive into the User Persona document. You'll
    find the link in your task materials. This document is your guide,
    revealing the user's background, their story, and exactly what they needed
    help with during the conversation.

- Embrace the Persona: Spend a few moments truly internalizing who this user
    is. Think about their situation, their potential emotions, and what would
    be most important to them. Step into their shoes and see the world from
    their perspective.
     [INFO] The data in the persona document is historical data and just to
         provide contexts. Therefore, the data involved in the conversations may
          not be exactly the same as the persona doc.

- Relive the Conversation: With your user hat firmly on, open the first
    Conversation Script (link also provided). Read the dialogue between the
    agent and the user (remember, that's you now!). As you go through it,
    filter everything through the lens of the user persona. How does each turn
    make you feel? Is the agent meeting your needs as described in the persona?

- Share Your (User) Experience: Finally, move on to the survey. Answer the
    questions based entirely on how you, playing the role of this specific user
    , experienced the interaction with the health agent.


Note 1: The same user independently talks to three agents (hense three
    conversation scripts). The user persona retains consistent in general
    across these conversations.

Note 2: Numeric accuracy is not the focus in this evaluation, so you don't
    need to worry about data details. It is possible that three agents may
    state slightly different values (e.g., different sleep durations) due to
    minor difference in computation methods. Please ignore such details and
    focus on the conversation flow.

Note 3: Formatting of the content is also not the focus in this evaluation.
    Please pay more attention to the actual content rather than the response
    format (e.g., bullet points, bolding).
```

Table S28 | PHA Generalist Raw Evaluation Questions after Reading Each Multi-Turn Conversation with Agents. Ranking questions are based on the same 18 rubrics after reading all conversations.

Question No.	Evaluation Question Body	Scoring/Ranking Question
Overall-1	If you were the user in this conversation, how satisfied are you after the conversation?	1: Very Dissatisfied - The conversation was unhelpful or frustrating. 2: Dissatisfied - The conversation had significant issues and did not meet my needs. 3: Neutral - The conversation was okay, but not particularly helpful or satisfying. 4: Satisfied - The conversation was generally helpful and met most of my needs. 5: Very Satisfied - The conversation was highly helpful, insightful, and fully met my needs.
Overall-2	What's the overall quality of the agent response?	1: Very Poor - Responses were consistently irrelevant, incorrect, confusing, or inappropriate. 2: Poor - Responses had frequent issues with relevance, correctness, clarity, or appropriateness. 3: Fair - Responses were a mix of good and bad; some were helpful, others were not. 4: Good - Responses were mostly relevant, correct, clear, and appropriate. 5: Excellent - Responses were consistently relevant, correct, clear, appropriate, and insightful.
DS-1	The agent involves all of my useful and relevant personal data to answer my questions. (Note: Data accessible to the agent: basic info, Fitbit, blood tests)	1: None of the relevant user data is referenced or used. 2: Only a few minor pieces of relevant user data are referenced; most important data is missed. 3: About half of the most relevant user data is referenced and used appropriately. 4: Most of the relevant and important user data is referenced and used appropriately. 5: All or nearly all relevant and important user data is referenced and used appropriately where applicable.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
DS-2	The agent does not involve irrelevant or not useful personal data.	1: Agent frequently uses irrelevant personal data, causing confusion or distraction. 2: Agent sometimes uses irrelevant personal data. 3: Agent occasionally uses slightly irrelevant data, but it doesn't significantly detract. 4: Agent very rarely uses irrelevant data; responses focus on pertinent information. 5: Agent exclusively uses relevant personal data; no irrelevant data is included.
DE-1	The agent involves useful and relevant health knowledge in the conversation to help answer my questions.	1: No relevant health knowledge is apparent in the agent's responses. 2: Very little relevant health knowledge is used, or it's applied incorrectly. 3: Some relevant health knowledge is used, but significant gaps exist or it's not always applied effectively. 4: Most responses appropriately incorporate relevant health knowledge. 5: Responses consistently and effectively integrate useful and relevant health knowledge.
DE-2	The agent does not involve irrelevant or not useful health domain knowledge.	1: Agent frequently includes irrelevant or incorrect health information, causing confusion. 2: Agent sometimes includes irrelevant health information or jargon. 3: Agent occasionally includes minor pieces of irrelevant knowledge, but it's not disruptive. 4: Agent very rarely includes irrelevant knowledge; focus remains on pertinent information. 5: Agent exclusively uses relevant and accurate health knowledge; no irrelevant information included.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
DS-DE-1	The agent provides useful interpretation and contextualization of my health data that is helpful for addressing my needs.	1: No interpretation provided, or interpretation is incorrect/useless. 2: Minimal interpretation provided, or it lacks context/utility. 3: Some useful interpretation and contextualization provided, but could be clearer or more insightful. 4: Good interpretation and contextualization provided, generally helpful. 5: Excellent, insightful interpretation and contextualization provided, very helpful for understanding my health status and needs.
HC-1	The agent's tone was supportive and encouraging.	1: Agent felt dismissive, critical, or completely unsupportive. 2: Agent offered minimal or superficial support/encouragement. 3: Agent was somewhat supportive and encouraging. 4: Agent was generally supportive and encouraging throughout the conversation. 5: Agent was consistently very supportive and encouraging, making me feel motivated.
HC-2	The agent understood my preferences and barriers.	1: Agent completely ignored or misunderstood my stated preferences and barriers, or did not about ask my context before providing suggestions. 2: Agent showed little understanding or acknowledgment of my preferences and barriers. 3: Agent showed some understanding but missed key aspects or didn't integrate them well. 4: Agent showed good understanding and consideration of most of my preferences and barriers. 5: Agent demonstrated a clear and accurate understanding of my preferences and barriers, incorporating them into the conversation.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-3	The agent handled conversation flow naturally and effectively.	1: Conversation flow felt very unnatural, robotic, or difficult to continue. 2: Conversation flow felt awkward or stilted at times. 3: Conversation flow was acceptable but could have been smoother or more logical. 4: Conversation flow was mostly natural, logical, and easy to follow. 5: Conversation flow was very natural, smooth, engaging, and effective.
HC-4	The agent responses were useful, appropriate and informative.	1: Responses were generally useless, inappropriate, or uninformative. 2: Responses were often not very useful, appropriate, or informative. 3: Responses were somewhat useful, appropriate, and informative, but inconsistent. 4: Responses were mostly useful, appropriate, and informative. 5: Responses were consistently highly useful, appropriate, and informative.
HC-5	The agent provides personalized suggestions at an appropriate time.	1: Agent provided no personalized suggestions, or suggestions felt completely ill-timed (too early, too late, random). 2: Agent provided suggestions, but they often felt poorly timed or generic. 3: Agent provided personalized suggestions at somewhat appropriate times, but the timing could be improved. 4: Agent generally provided personalized suggestions at appropriate and logical points in the conversation. 5: Agent provided personalized suggestions seamlessly and at highly opportune moments.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-6	The agent handled my feedback well.	1: Agent ignored, dismissed, argued with, or responded negatively to my feedback. 2: Agent acknowledged feedback minimally but didn't seem to understand or use it. 3: Agent acknowledged feedback but struggled to incorporate it or respond effectively. 4: Agent handled my feedback reasonably well, acknowledged it, and attempted to adjust. 5: Agent handled my feedback excellently, acknowledged it clearly, and adapted its responses or approach appropriately.
HC-7	The agent helped me talk about changing my behavior to better align with my goal.	1: Agent did not help me talk about behavior change at all, or hindered the discussion. 2: Agent made minimal or ineffective attempts to discuss behavior change related to my goal. 3: Agent initiated some discussion about behavior change, but it wasn't very deep or helpful. 4: Agent effectively guided the conversation towards discussing relevant behavior changes. 5: Agent skillfully and supportively helped me explore specific behavior changes needed to reach my goal.
HC-8	The agent helped me feel confident in my ability to achieve the goal.	1: The conversation undermined my confidence or made me feel less capable. 2: The conversation had little to no positive impact on my confidence. 3: The conversation slightly increased my confidence in achieving the goal. 4: The conversation noticeably helped me feel more confident about achieving the goal. 5: The conversation significantly boosted my confidence and belief in my ability to achieve the goal.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-9	The agent helped me align the goal with my priorities, preferences, and barriers.	1: Agent did not help connect the goal to my personal context (priorities, preferences, barriers). 2: Agent made minimal or unclear attempts to connect the goal to my context. 3: Agent helped somewhat in connecting the goal, but the alignment wasn't fully explored or convincing. 4: Agent effectively helped me see how the goal aligns with my priorities, preferences, and acknowledged barriers. 5: Agent skillfully helped me integrate the goal with my personal context, making it feel relevant and achievable despite barriers.
HC-DS-1	The agent's suggestion is individualized to my personal data.	1: Suggestions felt completely generic, showing no connection to my personal data. 2: Suggestions showed minimal or superficial connection to my personal data. 3: Suggestions were somewhat individualized but could have leveraged my data much more effectively. 4: Suggestions were clearly individualized and tailored based on my relevant personal data. 5: Suggestions were highly individualized, demonstrating insightful and effective use of my personal data.
HC-DE-1	The agent's suggestion is grounded on health knowledge	1: Suggestions seemed arbitrary or lacked any apparent basis in health knowledge; potentially unsafe. 2: Suggestions had a weak or unclear connection to established health knowledge or principles. 3: Suggestions were generally aligned with health knowledge but felt generic or superficial. 4: Suggestions were clearly grounded in relevant and sound health knowledge. 5: Suggestions were strongly grounded in relevant health knowledge, often explaining the rationale clearly.

Table S29 | PHA Generalist Evaluation Rubrics. Each rubric is an aggregation of one or multiple raw questions for better presentation.

Evaluation Rubrics	Question No.
Data Science Analysis (DS)	DS-1
Domain Knowledge Integration (DE)	DE-1, DE-2
Natural Conversation (HC)	HC-1, HC-4, HC-5
Personalized Coaching (HC)	HC-2, HC-3, HC-6, HC-9
User Empowerment (HC)	HC-7, HC-8
Health Data Interpretation (DS + DE)	DS-DE-1
Individualized Suggestions (DS + HC)	HC-DS-1
Grounded Suggestions (DE + HC)	HC-DE-1
Overall Quality (DS + DE + HC)	Overall-1

F.3.2. Task 2: Expert Evaluation Rubrics

Health experts receive the following instruction prior to reading the conversation scripts between a user persona (role-playing by an agent) with each personal health agent:

```
Your goal is to assess the conversation's quality from the experts' point of
    view. Get ready for the evaluation:

- Understand the Scenario: Review the User Persona document (link provided in
    your task materials). This context is crucial for understanding the user's
    situation, needs, and the challenges the agent faced. You need to grasp the
     user's background to accurately judge the agent's responses.
    [INFO] The data in the persona document is historical data and just to
        provide contexts. Therefore, the data involved in the conversations may
         not be exactly the same as the persona doc.

- Analyze the Agent's Performance: Now, open the first Conversation Script (
    link also provided). Read the dialogue, focusing critically on the agent's
    contributions. Analyze each of the agent's turns.

- Provide Your Expert Assessment: Finally, move on to the survey. Answer the
    questions based entirely on your professional evaluation of the agent's
    performance during this specific interaction.

Note 1: The same user independently talks to three agents (hense three
    conversation scripts). The user persona retains consistency in general
    across these conversations.

Note 2: Numeric accuracy is not the focus in this evaluation, so you don't
    need to worry about data details. It is possible that three agents may
    state slightly different values (e.g., different sleep durations) due to
    minor difference in computation methods. Please ignore such details and
    focus on the conversation flow.

Note 3: Formatting of the content is also not the focus in this evaluation.
    Please pay more attention to the actual content rather than the response
    format (e.g., bullet points, bolding).
```

Table S30 | **PHA Expert Raw Evaluation Questions after Reading Each Multi-Turn Conversation with Agents**. Ranking questions are based on the same 24 rubrics after reading all conversations.

Question No.	Evaluation Question Body	Scoring/Ranking Question
Overall-1	What's the overall quality of the agent response?	1: Very Poor Quality - Responses are clinically unsound, unsafe, irrelevant, or unusable. 2: Poor Quality - Responses frequently exhibit issues with clinical soundness, safety, or relevance. 3: Acceptable Quality - Responses are generally safe but may lack depth, relevance, or clinical nuance. 4: Good Quality - Responses are mostly sound, safe, relevant, and effective. 5: Excellent Quality - Responses are consistently sound, safe, relevant, effective, and demonstrate best practices.
Overall-2	How appropriate and effective was the overall conversation likely to be for a user with this profile and needs?	1: Highly Inappropriate/Ineffective - Conversation likely counterproductive, confusing, or potentially harmful for this user. 2: Inappropriate/Ineffective - Conversation poorly suited to the user's profile/needs; unlikely to be helpful. 3: Somewhat Appropriate/Effective - Conversation has some potential value but significant limitations or mismatches for this user. 4: Appropriate/Effective - Conversation is generally suitable for the user's profile/needs and likely helpful. 5: Highly Appropriate/Effective - Conversation is very well-suited to the user's profile/needs and likely highly beneficial.
DS-1	Did the agent appropriately utilize relevant user data that was available to it? Note: Data that is accessible to the agent: basic info (age, gender, height, weight), Fitbit records (step, sleep, activity), blood test results.	1: None/Minimal Utilization - Relevant available user data was ignored or barely used. 2: Poor Utilization - Key relevant data was frequently missed or clearly underutilized. 3: Fair Utilization - Some relevant data was used, but significant omissions or inconsistent application exist. 4: Good Utilization - Most relevant available user data was utilized appropriately. 5: Excellent Utilization - All relevant available user data was consistently and appropriately utilized.
DS-2	Did the agent appropriately avoid irrelevant or potentially noisy user data?	1: Very Poor Avoidance - Agent frequently focused on irrelevant/noisy data, potentially causing confusion or misdirection. 2: Poor Avoidance - Agent often included or emphasized irrelevant/noisy data. 3: Fair Avoidance - Agent sometimes included irrelevant/noisy data, but it didn't heavily derail the interaction. 4: Good Avoidance - Agent rarely included or focused on irrelevant/noisy data. 5: Excellent Avoidance - Agent consistently ignored irrelevant/noisy data, focusing only on pertinent information.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
DE-1	Was the health information and knowledge provided by the agent accurate and factual?	1: Very Inaccurate - Information provided contains major factual errors or contradicts established knowledge. 2: Inaccurate - Information contains significant factual errors. 3: Mixed Accuracy - Information contains some inaccuracies or lacks sufficient precision. 4: Mostly Accurate - Information is generally correct, potentially with very minor, non-critical inaccuracies. 5: Highly Accurate - Information consistently aligns with current, evidence-based scientific facts.
DE-2	Did the agent avoid presenting inaccurate or outdated, or potentially harmful health information?	1: Very Poor Avoidance - Agent presented information that is potentially harmful, clearly contraindicated, or dangerously inaccurate/outdated. 2: Poor Avoidance - Agent presented information that could be harmful or is significantly inaccurate/outdated. 3: Fair Avoidance - Agent avoided directly harmful information but may have included some inaccurate or outdated points. 4: Good Avoidance - Agent avoided harmful information and provided mostly accurate and current information. 5: Excellent Avoidance - Agent completely avoided presenting harmful, inaccurate, or outdated information.
DE-3	Was the health information and knowledge provided by the agent appropriate and relevant for the user's health status and context?	1: Highly Inappropriate/Irrelevant - Information provided was clearly mismatched to the user's context, goals, or health status. 2: Inappropriate/Irrelevant - Information was largely unsuitable or irrelevant for this user. 3: Somewhat Appropriate/Relevant - Information had some relevance but significant portions were generic or mismatched. 4: Appropriate/Relevant - Information was generally suitable and relevant to the user's context. 5: Highly Appropriate/Relevant - Information was specifically tailored and highly relevant to the user's context, goals, and health status.
DE-4	Did the agent avoid presenting irrelevant or potentially confusing health information?	1: Very Poor Avoidance - Agent frequently presented irrelevant information or information in a confusing way, hindering understanding. 2: Poor Avoidance - Agent often included irrelevant or confusing information. 3: Fair Avoidance - Agent sometimes included irrelevant or confusing information, but it was manageable. 4: Good Avoidance - Agent rarely included irrelevant or confusing information. 5: Excellent Avoidance - Agent consistently presented only relevant information clearly and concisely.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
DS-DE-1	Did the agent correctly interpret the available user data within its expected ranges and context?	1: Incorrect Interpretation - Interpretation was fundamentally flawed, wrong based on data, or ignored critical context. 2: Mostly Incorrect Interpretation - Contained significant errors in interpreting data values or trends. 3: Mixed Interpretation - Some correct interpretation mixed with noticeable errors or overlooked key contextual factors. 4: Mostly Correct Interpretation - Generally accurate interpretation with only minor issues, if any. 5: Correct & Contextualized Interpretation - Interpretation was accurate, sound, and considered relevant context.
DS-DE-2	Was the agent's interpretation and contextualization of the user's health data (integrating data points and knowledge) appropriate and useful?	1: Inappropriate/Useless - Interpretation lacked relevance, offered no practical value, or was potentially misleading. 2: Mostly Inappropriate/Useless - Interpretation offered little clinical value or practical application. 3: Somewhat Appropriate/Useful - Interpretation had some value but lacked depth, clear relevance, or practical application. 4: Appropriate/Useful - Interpretation was generally relevant, sound, and helpful. 5: Highly Appropriate/Useful - Interpretation was relevant, insightful, and provided clear, practical value for the user's understanding or action.
HC-1	Did the agent maintain an engaging and welcoming demeanor that fosters a positive and open environment?	1: Discouraging/Unwelcoming/Robotic - Tone felt negative, judgmental, dismissive, or overly impersonal, hindering openness. 2: Neutral/Indifferent - Tone lacked warmth or positive engagement; felt purely transactional. 3: Somewhat Engaging/Welcoming - Attempts engagement but felt inconsistent, superficial, or slightly awkward. 4: Engaging/Welcoming - Tone was generally positive, respectful, and fostered a good environment for communication. 5: Highly Engaging/Welcoming - Tone consistently fostered trust, rapport, openness, and positivity through warmth and respect.
HC-2	Did the agent show understanding, recognition, and encouragement in communication?	1: Dismissive/Unaware - Agent ignored or contradicted user's feelings/statements; lacked recognition or encouragement. 2: Minimal Recognition/Support - Very little evidence of understanding user's perspective; encouragement absent or generic. 3: Some Recognition/Support - Showed basic understanding or offered simple encouragement, but lacked depth or personalization. 4: Good Understanding/Support - Generally acknowledged user's perspective appropriately and offered suitable encouragement. 5: Highly Empathetic/Supportive - Consistently showed understanding (e.g., through reflections), validated user's perspective, and offered genuine, tailored encouragement.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-3	Did the agent enable the user to take charge of their actions and decisions confidently?	1: Disempowering/Prescriptive - Agent dictated actions, undermined user confidence, or overly controlled the conversation. 2: Little Empowerment - Agent offered little support for user autonomy or decision-making; interaction felt passive for the user. 3: Some Empowerment - Some effort to support user agency (e.g., offering choices), but inconsistent or weak. 4: Empowering - Agent generally encouraged user control, decision-making, and problem-solving. 5: Highly Empowering - Agent consistently used techniques (e.g., collaborative goal setting, exploring options) to support user autonomy and build self-efficacy.
HC-4	Did the agent's response concise, clear, and easy to follow, avoiding unnecessary jargon?	1: Confusing/Verbose/Jargony - Responses were consistently difficult to understand, overly long, or used inappropriate technical terms. 2: Often Unclear/Verbose - Responses were frequently difficult to follow, too wordy, or contained confusing jargon. 3: Moderately Clear - Responses were generally understandable but could often be improved for clarity, conciseness, or simpler language. 4: Clear/Concise - Responses were mostly easy to understand, appropriately detailed, and used suitable language. 5: Highly Concise/Clear - Responses were consistently easy to understand, succinct, well-structured, and used accessible, jargon-free language.
HC-5	Did the agent adequately elicit and appropriately incorporate the user's stated preferences and barriers?	1: Ignored/Dismissed - Agent failed to ask about or completely ignored stated preferences/barriers. 2: Poor Elicitation/Integration - Agent asked minimally or failed to use the elicited information effectively in suggestions or planning. 3: Fair Elicitation/Integration - Agent elicited some information, but its incorporation into the conversation or suggestions felt superficial or inconsistent. 4: Good Elicitation/Integration - Agent adequately elicited preferences/barriers and generally incorporated them appropriately. 5: Excellent Elicitation/Integration - Agent skillfully elicited preferences/barriers and thoughtfully integrated them into the conversation, goals, and suggestions.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-6	Was the timing of any personalized suggestions appropriate given the conversation flow?	1: Highly Inappropriate Timing - Suggestions felt random, premature, jarringly late, or interrupted the user inappropriately. 2: Poor Timing - Suggestions often felt ill-timed relative to the conversational context or user readiness. 3: Acceptable Timing - Suggestion timing was passable but could have been better aligned with the flow or user cues. 4: Good Timing - Suggestions were generally offered at logical and appropriate points in the conversation. 5: Excellent Timing - Suggestions were offered seamlessly and at highly opportune moments, feeling like a natural progression of the conversation.
HC-7	Did the agent respond appropriately and constructively to user feedback and questions?	1: Poor Response - Agent ignored, dismissed, argued with, or responded inaccurately/negatively to feedback or questions. 2: Weak Response - Agent acknowledged feedback/questions minimally but failed to address them effectively or accurately. 3: Fair Response - Agent acknowledged feedback/questions but struggled to respond effectively, accurately, or constructively. 4: Good Response - Agent responded appropriately to feedback-/questions and attempted to address them constructively. 5: Excellent Response - Agent responded accurately, respectfully, and constructively, effectively addressing or incorporating user feedback/questions.
HC-8	Did the agent appropriately guide the user in setting/refining goals that align with health principles and appear SMART (Specific, Measurable, Achievable, Relevant, Time-bound) within the user's context?	1: No Guidance/Poor Guidance - Agent failed to guide goal setting, guided towards inappropriate goals, or undermined the process. 2: Minimal/Ineffective Guidance - Agent's attempts at goal-setting guidance were weak, unclear, or did not lead to SMART or contextually relevant goals. 3: Some Guidance - Agent provided basic guidance but may have missed key SMART elements, relevance checks, or alignment with user context. 4: Good Guidance - Agent effectively guided the user towards setting/refining appropriate goals that meet most SMART criteria and fit the context. 5: Excellent Guidance - Agent skillfully and collaboratively guided the user to define/refine appropriate, personally meaningful SMART goals fully aligned with health principles and user context.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
HC-DS-1	Did the health information, advice, or coaching suggestions provided seem appropriately informed by the user's available personal health data (e.g., Fitbit, labs)?	1: No Connection Evident - Suggestions/advice seemed generic and completely unrelated to the user's specific data. 2: Weak Connection - Minimal or superficial link apparent between user data and suggestions/advice. 3: Moderate Connection - Some link between data and suggestions/advice, but data could have been leveraged more effectively or relevantly. 4: Good Connection - Suggestions/advice were generally well-informed by relevant aspects of the user's data. 5: Strong & Appropriate Connection - Suggestions/advice were clearly and appropriately tailored based on relevant user data insights.
HC-DE-1	Were the evidence-based health recommendations or explanations well-integrated within the coaching dialogue and behavior change suggestions?	1: No Integration/Contradictory - Health knowledge felt separate from coaching, or even contradicted it. 2: Poor Integration - Integration was weak, awkward; knowledge didn't clearly support or align with coaching elements. 3: Fair Integration - Some attempt at integration, but the connection between knowledge and coaching could be clearer or more effective. 4: Good Integration - Health knowledge and coaching elements were generally well-linked and mutually supportive. 5: Seamless Integration - Health knowledge was expertly woven into the coaching dialogue, explanations, and behavior change suggestions, providing clear rationale and support.
DS-DE-HC-1	Did the agent's responses feel coherent and internally consistent? Or were there parts of a response, or different responses, that seemed contradictory or disconnected from each other?	1: Highly Incoherent/Inconsistent - Responses frequently contained contradictions or felt disjointed and illogical. 2: Often Incoherent/Inconsistent - Noticeable contradictions or disjointedness occurred in several responses. 3: Somewhat Incoherent/Inconsistent - Occasional minor contradictions or a lack of smooth flow between ideas/components. 4: Mostly Coherent/Consistent - Responses were generally unified, logical, and flowed reasonably well. 5: Highly Coherent/Consistent - Responses consistently felt logical, unified, and flowed smoothly, integrating different elements seamlessly.

Continued on next page

Question No.	Evaluation Question Body	Scoring/Ranking Question
DS-DE-HC-2	Did the agent effectively synthesize user data, general health knowledge, and coaching strategies into unified and meaningful responses? Or did responses sometimes feel like a collection of separate, unlinked pieces of information?	1: No Synthesis - Information (data, knowledge, coaching) was presented as isolated, disconnected pieces. 2: Poor Synthesis - Very little connection or integration between different types of information or strategies. 3: Fair Synthesis - Some attempt at synthesis, but connections were often weak, incomplete, or not clearly articulated. 4: Good Synthesis - Different types of relevant information and strategies were generally combined effectively into meaningful responses. 5: Excellent Synthesis - Data, knowledge, and coaching strategies were skillfully integrated into cohesive, unified, and highly meaningful responses.
DS-DE-HC-3	Did the agent fail to identify or respond appropriately to any potential red flags or concerning user statements/data?	1: Critical Failure - Agent missed clear red flags requiring urgent/specific attention or responded highly inappropriately. 2: Significant Failure - Agent missed concerning statements/data or responded inappropriately. 3: Inconsistent Identification/Response - Agent identified some potential issues but missed others, or the response protocol was inconsistently or weakly applied. 4: Adequate Identification/Response - Agent identified most potential flags/concerning statements and responded reasonably appropriately. 5: Excellent Identification/Response - Agent appropriately identified potential red flags/concerns and responded in a good manner.
DS-DE-HC-4	Did the agent miss opportunities to ask crucial clarifying questions that a human expert likely would have asked?	1: Missed Crucial Opportunities - Agent consistently failed to ask obvious or necessary clarifying questions, leading to potential misunderstandings or poor advice. 2: Missed Several Opportunities - Agent often proceeded with assumptions where clarification was clearly needed. 3: Missed Some Opportunities - Agent sometimes failed to ask relevant clarifying questions, potentially limiting effectiveness. 4: Asked Most Necessary Questions - Agent generally asked appropriate clarifying questions when ambiguity or lack of detail was present. 5: Excellent Probing - Agent asked pertinent and insightful clarifying questions whenever necessary, similar to how a human expert would ensure understanding.

Table S31 | **PHA Expert Evaluation Rubrics**. Each rubric is an aggregation of one or multiple raw questions for better presentation.

Evaluation Rubrics	Question No.
Overall Quality (DS + DE + HC)	Overall-1, Overall-2
Appropriate User Data (DS)	DS-1
Accurate Health Info (DE)	DE-1
Appropriate Health Info (DE)	DE-3
User Data Interpretation (DS + DE)	DS-DE-1, DS-DE-2
Understanding & Engagement (HC)	HC-1, HC-2
User Empowerment (HC)	HC-3
Conciseness & Clarity (HC)	HC-4
Appropriate Timing (HC)	HC-6
Feedback Incorporation (HC)	HC-7
Guides Goal Setting (HC)	HC-8
Data-Informed Suggestions (DS + HC)	HC+DS-1
Knowledge Integration (DE + HC)	HC+DE-1
Coherent & Consistent (DS + DE + HC)	DS-DE-HC-1
Effective Synthesis (DS + DE + HC)	DS-DE-HC-2
Red Flags Identification (DS + DE + HC)	DS-DE-HC-3
Asks Clarifying Questions (DS + DE + HC)	DS-DE-HC-4

F.4. PHA Evaluation Results

F.4.1. Task 1: Multi-agent Conversation Experience for End-Users

Table S32 | PHA Rubrics Evaluation Results by End-users (N=20). All three agents – single-agent PHIA (ICC(3,1)=0.890), parallel multi-agent (ICC(3,1)=0.812), and PHA (ICC(3,1)=0.825) receive excellent IRR. Performance was the average scores of rubrics across DS, DE, HC agents, and their combinations. GLMM with Gamma link function was fitted for statistical tests. Post-doc Tukey's HSD were applied for pairwise comparison.

Rubrics	PHIA Performance	Parallel Multi-Agent Performance	PHA Performance	β (Multi-PHIA)	β (PHA-PHIA)	p-value (Multi-PHIA)	p-value (PHA-PHIA)	p-value (PHA-Multi)
Data Science Analysis (DS)	59.7±0.8 %	54.9±0.8 %	**70.3±0.7 %**	-0.080	0.175	<0.001***	<0.001***	<0.001***
Domain Knowledge Integration (DE)	84.9±0.3 %	84.9±0.3 %	**86.7±0.3 %**	0.000	0.022	0.995	<0.001***	<0.001***
Natural Conversation (HC)	83.7±0.3 %	83.6±0.3 %	**86.4±0.3 %**	-0.001	0.032	0.953	<0.001***	<0.001***
Personalized Coaching (HC)	83.4±0.3 %	83.7±0.3 %	**86.2±0.3 %**	0.004	0.034	0.64	<0.001***	<0.001***
User Empowerment (HC)	84.1±0.3 %	84.0±0.3 %	**86.3±0.3 %**	-0.001	0.026	0.967	<0.001***	<0.001***
Health Data Interpretation (DS + DE)	69.1±0.8 %	64.8±0.8 %	**76.5±0.6 %**	-0.068	0.114	<0.001***	<0.001***	<0.001***
Individualized Suggestions (DS + HC)	74.1±0.6 %	73.7±0.6 %	**79.2±0.5 %**	-0.008	0.074	0.819	<0.001***	<0.001***
Grounded Suggestions (DE + HC)	84.7±0.4 %	84.7±0.4 %	**86.3±0.3 %**	-0.001	0.020	0.991	<0.001***	<0.001***
Overall Quality (DS + DE + HC)	82.3±0.3 %	82.7±0.3 %	**85.1±0.4 %**	0.005	0.033	0.572	<0.001***	<0.001***

Table S33 | PHA Side-by-Side Comparison Results by End-users (N=20), with moderate IRR (Fleiss' κ =0.468). Performance was measured as the ranking rate. Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Friedman tests were applied for statistical tests, followed by post-hoc Wilcoxon Rank-sum tests with Holm-Bonferroni correction.

Rubrics	PHIA Ranking	Parallel Multi-Agent Ranking	PHA Ranking	Friedman Stats	p-value (Multi-PHIA)	p-value (PHA-PHIA)	p-value (PHA-Multi)
Preference	#1: 26.5%±1.4% #2: 32.8%±1.5% #3: 40.7%±1.5%	#1: 24.8%±1.4% #2: 36.7%±1.5% #3: 38.6%±1.6%	**#1: 48.7%±1.5%** **#2: 30.5%±1.5%** **#3: 20.8%±1.3%**	χ^2=116.77 p<0.001***	0.92	<0.001***	<0.001***
Response Quality	#1: 26.2%±1.4% #2: 33.4%±1.5% #3: 40.4%±1.6%	#1: 24.6%±1.4% #2: 36.7%±1.6% #3: 38.7%±1.6%	**#1: 49.1%±1.6%** **#2: 29.9%±1.5%** **#3: 21.0%±1.3%**	χ^2=118.44 p<0.001***	0.802	<0.001***	<0.001***

F.4.2. Task 2: Multi-agent Conversation Experience for Experts

Table S34 | PHA Rubrics Evaluation Results by Experts (N=5). All three agents – single-agent PHIA (ICC(3,1)=0.630), parallel multi-agent (ICC(3,1)=0.545), and PHA (ICC(3,1)=0.690) receive moderate to good IRR. Performance was the average scores of rubrics across DS, DE, HC agents, and their combinations. GLMM with Gamma link function was fitted for statistical tests. Post-doc Tukey's HSD were applied for pairwise comparison.

Rubrics	PHIA Performance	Parallel Multi-Agent Performance	PHA Performance	β (Multi-PHIA)	β (PHA-PHIA)	p-value (Multi-PHIA)	p-value (PHA-PHIA)	p-value (PHA-Multi)
Appropriate User Data (DS)	44.9±1.5 %	46.9±1.6 %	**69.4±1.4 %**	0.050	0.482	0.42	<0.001***	<0.001***
Accurate Health Info (DE)	81.8±0.9 %	81.1±1.0 %	**87.4±0.8 %**	-0.009	0.069	0.8	<0.001***	<0.001***
Appropriate Health Info (DE)	72.6±1.0 %	75.0±1.1 %	**85.5±0.9 %**	0.034	0.171	0.119	<0.001***	<0.001***
Appropriate Timing (HC)	65.4±1.3 %	68.2±1.2 %	**85.4±0.8 %**	0.059	0.313	0.037*	<0.001***	<0.001***
Asks Clarifying Questions (HC)	38.6±1.1 %	49.4±1.2 %	**77.6±1.2 %**	0.265	0.75	<0.001***	<0.001***	<0.001***
Feedback Incorporation (HC)	74.0±0.7 %	76.6±0.8 %	**85.7±0.8 %**	0.033	0.146	0.057	<0.001***	<0.001***
Understanding & Engagement (HC)	55.4±0.9 %	66.4±1.0 %	**86.1±0.8 %**	0.185	0.458	<0.001***	<0.001***	<0.001***
Conciseness & Clarity (HC)	70.2±1.0 %	65.0±1.1 %	**81.1±0.9 %**	-0.081	0.149	<0.001***	<0.001***	<0.001***
User Empowerment (HC)	48.9±1.0 %	59.2±1.2 %	**81.2±1.0 %**	0.195	0.534	<0.001***	<0.001***	<0.001***
Guides Goal Setting (HC)	40.6±0.7 %	44.8±0.9 %	**63.1±1.2 %**	0.095	0.431	<0.001***	<0.001***	<0.001***
User Data Interpretation (DS + DE)	62.7±1.4 %	63.9±1.4 %	**77.2±1.2 %**	0.013	0.241	0.875	<0.001***	<0.001***
Data-Informed Suggestions (DS + HC)	47.5±1.3 %	50.0±1.4 %	**69.7±1.4 %**	0.047	0.4	<0.001***	<0.001***	<0.001***
Knowledge Integration (DE + HC)	63.8±1.0 %	67.4±1.0 %	**81.8±0.8 %**	0.061	0.262	0.005**	<0.001***	<0.001***
Overall Quality (DS + DE + HC)	65.2±0.8 %	70.8±0.9 %	**84.8±0.9 %**	0.083	0.265	<0.001***	<0.001***	<0.001***
Effective Synthesis (DS + DE + HC)	60.7±1.1 %	66.2±1.0 %	**80.4±0.9 %**	0.093	0.296	<0.001***	<0.001***	<0.001***
Coherent & Consistent (DS + DE + HC)	77.8±0.6 %	76.6±0.7 %	**85.4±0.8 %**	-0.016	0.092	0.405	<0.001***	<0.001***
Red Flags Identification (DS + DE + HC)	73.5±1.0 %	75.7±0.9 %	**81.4±1.0 %**	0.031	0.103	0.283	<0.001***	0.001**

Table S35 | PHA Side-by-Side Comparison Results by Experts (N=5). Our results showed excellent agreement rate (81.5%). Performance was measured as the ranking rate. Bootstrap was applied (1000 times sampling with replacement) to calculate the standard error. Friedman tests were applied for statistical tests, followed by post-hoc Wilcoxon rank-sum tests with Holm-Bonferroni correction.

Rubrics	PHIA Ranking	Parallel Multi-Agent Ranking	PHA Ranking	Friedman Stats	p-value (Multi-PHIA)	p-value (PHA-PHIA)	p-value (PHA-Multi)
Overall Quality	#1: 4.4%±1.3% #2: 28.8%±2.9% #3: 66.8%±3.0%	#1: 15.6%±2.3% #2: 59.3%±3.0% #3: 25.2%±2.7%	**#1: 80.0%±2.5%** **#2: 12.1%±2.0%** **#3: 8.0%±1.7%**	χ^2=229.25 p<0.001***	<0.001***	<0.001***	<0.001***
Effectiveness Usefulness	#1: 4.8%±1.4% #2: 30.3%±2.9% #3: 64.8%±3.0%	#1: 15.2%±2.3% #2: 58.8%±3.0% #3: 25.9%±2.7%	**#1: 80.0%±2.5%** **#2: 10.8%±1.9%** **#3: 9.2%±1.9%**	χ^2=218.23 p<0.001***	<0.001***	<0.001***	<0.001***