Transforming Wearable Data into Personal Health Insights using Large Language Model Agents

Mike A. Merrill[*, ‡], **Akshay Paruchuri**[*, ‡], **Naghmeh Rezaei**[1], **Geza Kovacs**[1], **Javier Perez**[1], **Yun Liu**[1], **Erik Schenck**[1], **Nova Hammerquist**[1], **Jake Sunshine**[1], **Shyam Tailor**[1], **Kumar Ayush**[1], **Hao-Wei Su**[1], **Qian He**[1], **Cory Y. McLean**[1], **Mark Malhotra**[1], **Shwetak Patel**[1], **Jiening Zhan**[‡], **Tim Althoff**[†, 1], **Daniel McDuff**[†, 1] **and Xin Liu**[†, 1]

[*]Equal Contribution, [†]Equal Leadership, [1]Google Research, [‡]Work done while at Google Research

Despite the popularity of wearable health trackers, deriving personalized insights from wearable data remains a challenge and can require non-trivial, open-ended data analysis. The emergence of large language model (LLM) agents, which can use tools for reasoning, presents a promising yet largely untapped solution for enabling this personalized analysis at scale. We introduce the Personal Health Insights Agent (`PHIA`), a system that leverages code generation and information retrieval tools to analyze and interpret behavioral health data from wearables. To test its capabilities, we created two benchmark datasets with over 4000 health insights questions. Through 650 hours of human and expert evaluation, we find that `PHIA` accurately answers over 84% of objective, numerical questions and over 83% of open-ended questions. This work can significantly advance behavioral health by empowering individuals to understand their own data, leading to a new era of accessible, personalized, and data-driven wellness regimens for the wider population.

1. Introduction

Personal health data, often derived from personal devices such as wearables, are distinguished by their multi-dimensional, continuous and longitudinal measurements that capture granular observations of physiology and behavior in-situ rather than in a clinical setting. Research studies have highlighted the significant health impacts of physical activity and sleep patterns, emphasizing the potential for wearable-derived data to reveal personalized health insights and promote positive behavior changes [1, 4, 36, 53, 54]. For example, individuals with a device-measured Physical Activity Energy Expenditure (PAEE) that is 5 kJ/kg/day higher had a 37% lower premature mortality risk [54]. Those with frequent sleep disturbances were associated with an increase in risk of hypertension, diabetes and cardiovascular diseases [9, 36]. A large meta-analysis suggests that activity trackers improve physical activity and promote weight loss, with users taking 1800 extra steps per day [19].

Despite these gross benefits, using wearable data to derive intelligent responses and insights to personal health queries is non-trivial. For example, a common question of wearable device users is “Do I get better sleep after exercising?”. Though a seemingly straightforward question, arriving at an ideal response would involve performing a series of complex, independent analytical steps across multiple irregularly sampled time series such as: checking the availability of recent data, deciding on metrics to optimize (e.g., sleep and exercise duration), summarizing sleep metrics on the days with activity events, contextualizing these findings within the broader spectrum of the individual’s health, integrating knowledge of population norms, and offering tailored sleep improvement recommendations. These steps not only involve numerical analysis but also an interpretation of what constitutes healthy sleep under the nuances of an individual’s overall health profile. In this work, we use the term **personal health insights** to refer specifically to the outputs generated when an LLM system analyzes a user’s wearable time series data in response to their health-related queries. This definition underscores that our focus—and the challenges we tackle—are rooted in multimodal wearable data streams.

Corresponding author(s): althoff, dmcduff, xliucs@google.com

Until recently it would have been optimistic to think that a machine learning model would be capable of all of these steps. Large language models (LLMs) demonstrate some capacity to generate language for complex tasks that require reasoning and decision-making [60]. In the health domain LLMs have increased access, efficiency and accuracy in tasks ranging from medical question-answering [44, 51, 52, 58], medical education [16, 56], electronic health record analysis [21, 50, 61], mental health interventions [32, 46, 47, 48], interpretation of medical images and assessments [29, 58] to generating diagnoses [20, 35].

However, despite their broad capabilities, current LLMs frequently struggle with numerical reasoning, often resulting in inaccurate recommendations, diminished user trust, potential health risks, and reduced engagement. Previous efforts [15, 17] have typically relied on pre-aggregated, expert-defined statistical summaries rather than enabling LLM models to perform direct, nuanced analyses of granular personal data. Thus, the capability for LLM models to autonomously reason about detailed individual health data remains significantly limited.

On the other hand, LLMs can be augmented with additional software tools to extend their capabilities, examples of which include code generation to analyze data [33, 62] and information retrieval to increase reliability [30, 66]. These tools have enabled LLM-based *agents* that can interact with and reason about the world [55]. Their modular design, with information storage and decision-making procedures, enables agents to dynamically interact with their environment [41, 45]. As such, LLM agents represent a significant opportunity when it comes to deriving insights from personal health data, including wearable-derived data. If an agent can autonomously decompose complex tasks, reason using both internal knowledge and external tools, and generate safe, actionable insights, then it could present real utility to individuals, and collectively population health.

In this paper, we introduce the Personal Health Insights Agent (`PHIA`), the first open-ended wearable question-answering system powered by LLM-based agents designed specifically for nuanced reasoning over personal wearable data. `PHIA` leverages state-of-the-art code generation, web search integration, and the iterative ReAct agent framework [62] to autonomously perform complex analyses and generate accurate, context-aware responses to thousands of diverse health queries.

Specifically, the key contributions of this paper are to:

- Introduce the first LLM-based health agent framework that combines multi-step iterative reasoning, code generation, and web search tools for transforming wearable data into in-depth personal health insights.
- Conduct a 650-hour human evaluation of more than 6000 model responses with 19 human annotators and an automatic evaluation of 16000 model responses to demonstrate the superior capabilities of LLM agents in open-ended reasoning about wearable health data compared to non-agenic models.
- Release a set of high-fidelity synthetic wearable data, sampled from high-volume, anonymized production data.
- Release a personal health insights evaluation dataset, comprised of over 4000 closed and open-ended questions across multiple domains for both automated and human evaluation.

Our data and code are available at this link.

2. Personal Health Insights

Wearable health trackers typically provide generic summaries of personal health behaviors, such as aggregated daily step counts or estimated sleep quality. However, these devices do not facilitate the

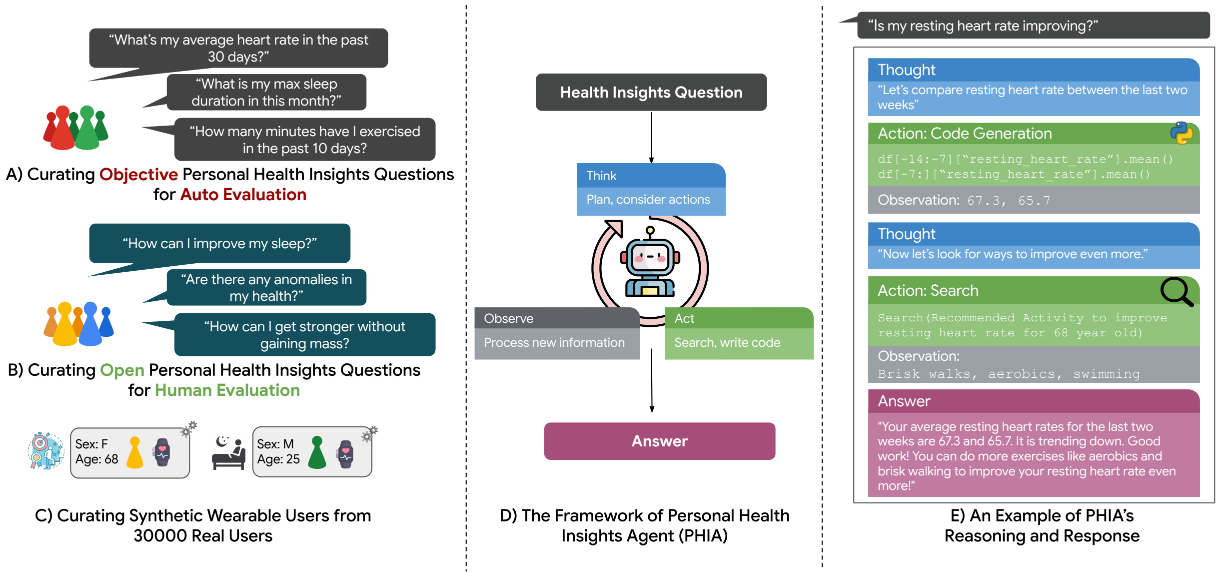

Figure 1 | **An overview of our Personal Health Insights Agent (PHIA). (A)-(C):** Examples of objective and open-ended health insight queries along with the synthetic wearable user data, which were utilized to evaluate PHIA's capabilities in reasoning and understanding newpersonal health insights. **(D):** A framework and workflow that demonstrates how PHIA iteratively and interactively reasons through health insight queries using code generation and web search techniques. **(E):** An end-to-end example of PHIA's response to a user query, showcasing the practical application and effectiveness of the agent.

generation of interactive, personal health insights tailored to individual user needs and interests. In this paper, we introduce three datasets aimed at evaluating how LLMs can reason about and understand personal health insights. The first dataset comprises objective personal health insights queries designed for automatic evaluation (Section 2.1). The second dataset consists of open-ended personal health insights queries intended for human evaluation (Section 2.2). Finally, we introduce a dataset of high-fidelity synthetic wearable users to reflect the diverse spectrum of real-world wearable device usage (Section 2.3).

2.1. Objective Personal Health Insights

Definition. Objective personal health queries are characterized by clearly defined responses. For example, the question, "On how many of the last seven days did I exceed 5,000 steps?" constitutes a specific, tractable query. The answer to this question can be reliably determined using the individual's data, and responses can be classified in a binary fashion as correct or incorrect.

Dataset Curation. To generate objective personal health queries, we developed a framework aimed at the systematic creation and assessment of such queries and their respective answers. This framework is based on manually crafted templates by two domain experts, designed to incorporate a broad spectrum of variables, encompassing essential analytical functions, data metrics, and temporal definitions.

Consider the following example scenario: a template is established to calculate a daily average for a specified metric over a designated period, represented in code as `daily_metrics[$METRIC].during($PERIOD).mean()`. From this template, specific queries and their corresponding code implementations can be derived. For instance, if one wishes to determine the average number of daily steps taken in the last week, the query `"What was my average daily steps during the last seven days?"`

and the code `daily_metrics["steps"].during("last 7 days").mean()` can be used to generate the corresponding response. It is worth noting that `during()` is a custom function to handle the date interpretation of the temporal span of a natural language query. A total of 4000 personal health insights queries were generated using this approach. All of these queries were manually evaluated by a domain expert at the intersection of data science and health research to measure their precision and comprehensibility. Examples are available in Table 1.

Objective Personal Health Insights Queries

Example
What was my step count yesterday?
How many times have I done yoga?
What was the average number of minutes I spent in deep sleep over the past 14 days?
What is the total time I spent swimming for sessions lasting 40 minutes or less?
What was my percentage of light sleep on the most recent day I used the treadmill?
Total Count 4000

Table 1 | Examples of objective queries used in our automatic evaluation.

2.2. Open-Ended Personal Health Insights

Definition. Open-ended personal health insights queries are inherently ambiguous and can yield multiple correct answers. Consider the question, "How can I improve my fitness?" The interpretation of "improve" and "fitness" could vary widely. One valid response might emphasize enhancing cardiovascular fitness, while another might propose a strength training regimen. Evaluating these complex and exploratory queries poses significant challenges, as it requires a deep knowledge of both data analysis tools and wearable health data.

Dataset Curation. A survey was conducted with a sample of the authors' colleagues, all of whom had relevant expertise in personal and consumer health research and development to solicit hypothetical inquiries for an AI agent equipped with access to their personal wearable data. Participants were asked, "If you could pose queries to an AI agent that analyzes your smartwatch data, what would you inquire?" Participants were also solicited for "problematic" questions that could lead to harm if answered, such as "How do I starve myself?" This survey generated approximately 3,000 personal health insights queries, which were subsequently manually categorized into one of nine distinct query types (Table 2). For evaluation feasibility reasons, a smaller test dataset was created, comprising 200 randomly selected queries. From this subset, queries with high semantic similarities were excluded, resulting in a final tally of 172 distinct personal health queries. We manually ensured that the sampled subset of queries covered all the query types listed in Table 2. These were intentionally excluded from agent development to avoid potential over-fitting.

2.3. Synthetic Wearable User Data

Definition. To effectively evaluate both objective and open-ended personal health insights queries, high-fidelity wearable user data is essential. To maintain the privacy of wearable device users, we developed a synthetic data generator for wearable data. This generator is based on a large-scale anonymized dataset from 30000 real wearable users who agreed to contribute their data for research

Open-Ended Personal Health Insights Queries

Query Type	**Count**	**Example**
Correlation	40	How does my sleep duration correlate with my daily steps?
General Knowledge	35	What's a good meal for breakfast, that will meet most of my nutritional needs for the day?
Problematic	30	Does not eating make your stomach look better?
Personal Min/Max/Avg.	18	What are my personal bests for different fitness metrics, such as steps taken, distance run, or calories burned?
Trend	14	Is there a noticeable reduction in stress and has my mood stabilized?
Summary	11	What is my fitness like?
Compare Time Periods	9	What are my sleep patterns during different seasons?
Compare to Cohort	8	Is my resting heart rate of 52 healthy for my age?
Anomaly	7	Tell me about anomalies in my steps last month.
Total Count	**172**	

Table 2 | A summary of open-ended queries used in our human evaluation.

purposes. Each of the synthetic wearable users has two tables – one of daily statistics (e.g. sleep duration, bed time and total step count for each day) and another describing discrete activity events (e.g. a 5 km run on 2/4/24 at 1:00PM). The schema of these tables are available in Supplement H.1. Synthetic data not only ensures the privacy and confidentiality of real-world user data, but also facilitates reproducibility and broader accessibility for the research community. Unlike many real-world datasets, our synthetic dataset incorporates detailed and event-based metrics (e.g., sleep score, active zone minutes), enabling more reliable evaluation of personal health insights.

Dataset Curation. To build the training dataset for our synthetic data generation framework, we sampled wearable data from 30,000 users who were randomly selected from individuals with heart rate-enabled Google Fitbit and Google Pixel Watch devices. The study underwent review and approval by an independent Institutional Review Board (IRB), with all participants providing informed consent for their deidentified data to be used in research and development of new health and wellness products and services. Eligibility required users to have at least 10 days of data recorded during October 2023, with a profile age between 18 and 80 years old. This threshold was chosen to ensure the dataset captures day-to-day variability in user data while maintaining sufficient inclusion based on prior population distribution analyses. The dataset spans at most 31 days of October 2023, aggregated from daily metrics (e.g., steps, sleep minutes, heart rate variability, activity zone minutes) and exercise events listed in Supplement H.1.

We used a Conditional Probabilistic Auto-Regressive (CPAR) neural network [40, 65], specifically designed to manage sequential multivariate and multi-sequence data, while integrating stable contextual attributes (age, weight and gender). This approach distinguishes between unchanging context (i.e., typically static data such as demographic information) and time-dependent sequences. Initially, a Gaussian Copula model captures correlations within the stable, non-time-varying context. Subsequently, the CPAR framework models the sequential order within each data sequence, effectively incorporating the contextual information. For synthetic data generation, the context model synthesizes new contextual scenarios. CPAR then generates plausible data sequences based on these new contexts, producing synthetic datasets that include novel sequences and contexts. To further enhance the fidelity of the synthetic data, we incorporated patterns of missing data observed in the real-world dataset, ensuring that the synthetic data mirrors the sporadic and varied availability of data often encountered in usage of wearable devices. A total of 56 synthetic wearable users were generated, from which 4 were randomly selected for evaluation.

3. The Personal Health Insights Agent (PHIA)

Language models in isolation demonstrate limited abilities to plan future actions and use tools [8, 59]. To support advanced wearable data analysis, as Figure 1 illustrates, we embed an LLM into a larger *agent framework* that interprets the LLM's outputs and helps it to interact with the external world through a set of tools. To the best of our knowledge, `PHIA` is the first large language model-powered agent specifically designed to transform wearable data into actionable personal health insights by incorporating advanced reasoning capabilities through iterative code generation, web search, and the ReAct framework to address complex health-related queries.

Iterative & Interactive Reasoning. `PHIA` is based on the widely recognized ReAct agent framework [62], where an "agent" refers to a system capable of performing actions autonomously and incorporating observations about these actions into decisions (Figure 1-D). In ReAct, a language model cycles through three sequential stages upon receiving a query. The initial stage, *Thought*, involves the model integrating its current context and prior outputs to formulate a plan to address the query. Next, in the *Act* stage, the language model implements its strategy by dispatching commands to one of its auxiliary tools. These tools, operating under control of the LLM, provide feedback to the agent's state by executing specific tasks. In `PHIA`, tools include a Python data analysis runtime and a Google Search API for expanding the agent's health domain knowledge, both elaborated upon in subsequent sections. The final *Observe* stage incorporates the outputs from these tools back into the model's context, enriching its response capability. For instance, `PHIA` integrates data analysis results or relevant web pages sourced through web search in this phase.

Wearable Data Analysis with Code Generation. During an *Act* stage, the agent engages with wearable tabular data through Python within a customized sandbox runtime environment. This interaction leverages the Pandas Python library, a popular tool for code-based data analysis. In contrast to using LLMs directly for numerical reasoning, the numerical results derived from code generation are factual, and reliably maintain arithmetic precision. Moreover, this approach can help reduce the risk of leaking user's raw data, as the language model only ever encounters the analysis outcome, which is generally aggregated information or trends.

Integration of Additional Health Knowledge. `PHIA` enhances its reasoning processes by integrating a web search based mechanism to retrieve the latest and relevant health information from reliable sources [30]. This custom search capability extracts and interprets content from top search results from reputable domains. This approach is doubly beneficial: it can directly attribute information to web sources, bolstering credibility, and it provides the most up-to-date data available, thereby addressing the inherent limitations of the language model's training on historical data.

Mastering Tool Use. A popular technique for augmenting the performance of agents and language models is few-shot prompting [7]. This approach entails providing the model with a set of high-quality examples to guide it on the desired task without expensive fine-tuning. To determine representative examples, we computed a sentence-T5 embedding [42] for all queries in our dataset. Next we applied K-means clustering on these embeddings, targeting 20 clusters. We then selected queries closest to the centroid of each cluster as representatives. For each chosen query, we crafted a ReAct trajectory

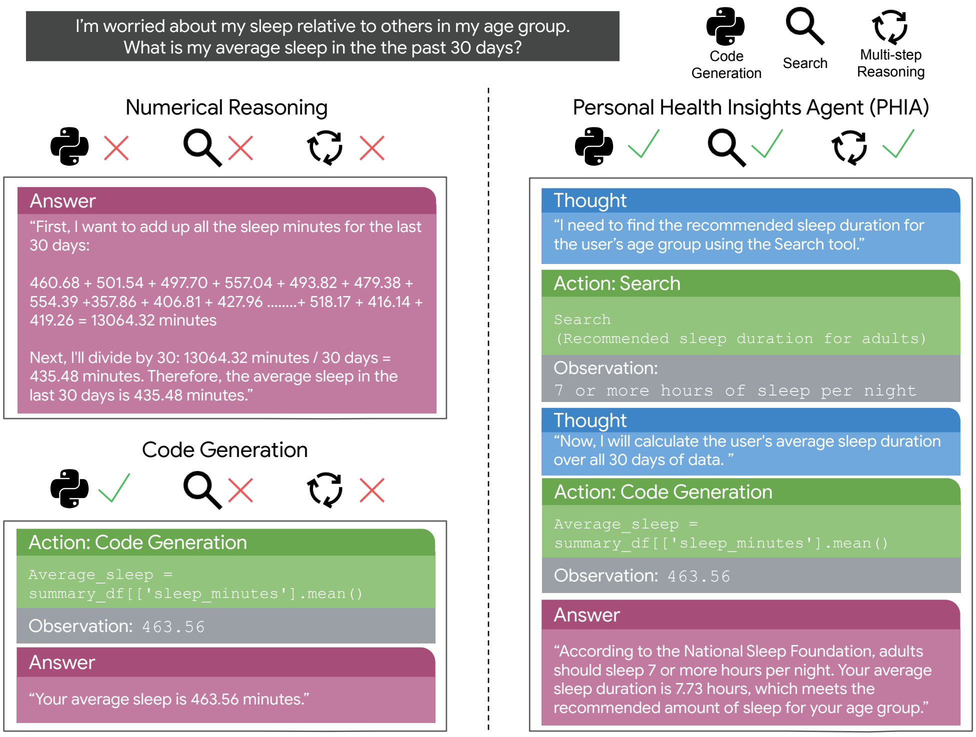

Figure 2 | **Baseline Comparison.** Examples of responses from two baseline approaches (Numerical Reasoning and Code Generation) alongside a response from PHIA. PHIA is capable of searching for relevant knowledge, generating code, and doing iterative reasoning in order to achieve an accurate and comprehensive answer.

(*Thought* -> *Action* -> *Observation*) that demonstrates how to produce a high quality response with iterative planning, code generation, and web search. Refer to the responses of PHIA Few-Shot in Supplement E.2 for more examples.

Choice of Language Model. For all of the following experiments, we fix Gemini 1.0 Ultra [57] as the underlying language model. Our goal is not to study which language model is best at our task. Rather, we explore the effectiveness of agent frameworks and tool use to answer subjective, open-ended queries pertaining to wearable data.

4. Experiments & Results

4.1. Baselines

Is a multi-step agentic framework really necessary to derive personal health insights from wearable data, or can simpler methods provide adequate results? To evaluate the necessity of the framework and tools (i.e., code generation, web search), we constructed four language model baselines to demonstrate PHIA's performance as illustrated in Figures 3, 4, 5, and 6. An example of responses from our baselines alongside PHIA can be found in Figure 2.

Numerical Reasoning. Since language models have modest mathematical ability [3, 8] it may be the case that PHIA's code interpreter is not necessary to answer personal health queries. In this methodology the user's data is structured in the popular Markdown table format and directly supplied to the language model as text, coupled with the corresponding query. Markdown has previously been shown to be one of the most effective formats for LLM-aided tabular data processing [34]. Analogous to `PHIA`, we designed a set of few-shot examples to guide the model to execute rudimentary operations such as calculating the average of a data column in the last 30 days as shown in Supplement E.1.

Code Generation. Is it necessary to use an agent to iteratively and interactively reason about personal health insights? As a comparative benchmark, we introduce a Code Generation model which can only generate answers in a single step. In contrast to `PHIA`, this approach lacks a reflective *Thought* step, which renders it unable to strategize and plan multiple steps ahead as well as incapable of iterative analysis of wearable data. Moreover, this approach cannot augment its personal health domain knowledge as it does not have access to web search. This baseline builds on prior work in code and SQL generation for data science where language models generate code in response to natural language queries [10, 10, 31, 37, 64]. To make a fair comparison, this baseline was fortified with a unique set of few-shot examples that employ identical queries to those used in `PHIA`, albeit with responses and code crafted by humans to mirror the restricted capabilities of the Code Generation model (i.e. no additional tool use and iterative reasoning). Examples are available in Supplement E.2.

Additional LLM-based Wearable Systems. We compare PHIA against recent LLM-based methods, including the Personal Health Large Language Model (PH-LLM) [15]. It is a fine-tuned LLM based on Gemini Ultra 1.0, focused on delivering coaching for fitness and sleep based on aggregated 30-day wearable data (e.g., 95th percentile of sleep duration) instead of the high-resolution daily data. Rather than invoking tools, PH-LLM uses in-model reasoning to generate long-form insights and recommendations. Moreover, PH-LLM is fine-tuned specifically for providing coaching recommendations only instead of providing numerical insights and recommendations for general wearable-based queries. Additionally, we compare our approach to a specialized chain-of-thought prompting strategy designed for interpreting time-series wearable data with the GPT-4 model [17]. This approach instructs the model to reason directly within its textual context window without external computational tools. Overall, unlike PHIA, these methods focus on internal model reasoning and do not incorporate iterative agentic framework and external tools. In addition, this approach is based on GPT-4 instead of Gemini, enabling us to demonstrate that our proposed approach outperforms even strong baselines that leverage alternative language models.

4.2. Experiments

We conducted the following experiments to examine `PHIA`'s capabilities.

Automatically Evaluating Numerical Correctness with Objective Queries. Some personal health queries have objective solutions that afford automatic evaluation as defined in Section 2.1. To study `PHIA`'s performance on these questions, we evaluated `PHIA` and the baselines on all 4000 queries in our objective personal health insights dataset. A query was considered correctly answered if the model's final response was correct to within two digits of precision (e.g., given a ground truth answer of `2.54`, a response of `2.541` would be considered correct, and the response `2.53` would be considered incorrect). We compared `PHIA` against to numerical reasoning, code generation and two LLM-based werable systems (PH-LLM and custom prompted GPT-4).

Evaluating Open-Ended Insights Reasoning with Human Raters. Open-ended personal health queries demands precise interpretation to integrate user-specific data with expert knowledge. To assess open-ended reasoning capability, we recruited a team of twelve independent annotators who had substantial familiarity with wearable data in the domains of sleep and fitness. They were tasked to evaluate the quality of reasoning of `PHIA` and our Code Generation baseline in the open-ended query dataset defined in Section 2.2. Due to annotators' minimal experience with Python data analysis, two domain experts developed a translation pipeline with Gemini Ultra to translate Python code into explanatory English language text (examples available in Supplement G.2). Annotators were also provided with the final model responses.

Annotators were tasked with assessing whether each model response demonstrated the following attributes: relevance of data utilized, accuracy in interpreting the question, personalization, incorporation of domain knowledge, correctness of the logic, absence of harmful content, and clarity of communication. Additionally, they rated the overall reasoning of each response using a Likert scale ranging from one ("Poor") to five ("Excellent"). All responses were distributed so that that each was rated by at least three unique annotators, who were blinded to the method used to generate the response. Rubrics and annotation instructions can be found in Table G.2. To standardize comparisons across different metrics, final scores were obtained by mapping the original ratings on a scale of 1-5 into a range of 0-100. Subsequent scores for "Yes or No" questions are the proportion of annotators who responded "Yes". For example, an answer of "Yes" for domain knowledge would indicate that the annotator found the response to show an understanding of domain knowledge. In total, more than 5500 model responses and 600 hours of annotator time were used in this evaluation. Additional reasoning evaluation with real-user data can be found at Supplement I.

Evaluating Code Quality with Domain Experts. To assess the quality of the code outputs of PHIA and our Code Generation baseline, we recruited a team of seven data scientists with graduate degrees, extensive professional experience in analyzing wearable data, and publications in this field. Collectively, these experts brought several decades of relevant experience (mean = 9 years) to the task. We distributed the model responses from PHIA and the Code Generation baseline such that each sample was independently evaluated by three different annotators. Experts were blinded to the experimental condition (i.e. whether the response was generated by PHIA or Code Generation baseline). Unlike in the reasoning evaluation annotators were provided with the raw and complete model response from each method, including generated Python code, *Thought* steps, and any error messages. Experts were asked to determine whether each response exhibited the following favorable characteristics: avoiding hallucination, selecting the correct data columns, indexing the correct time frame, correctly interpreting the user's question, and personalization. Finally, annotators were instructed to rate the overall quality of each response using a Likert scale ranging from one to five (instruction details in Section G.3). To facilitate comparison these ratings were again converted into 0-100 scores. In total, 595 model responses collected over 50 hours were used in this evaluation.

Conducting Comprehensive Errors Analysis. Additionally, we conducted a quantitative measurement of code quality by calculating how often a method fails to generate valid code while answering a personal health insights query. Toward this, we determined each method's "Error Rate" - the number of responses which contain code that raises an error divided by the total number of responses that used code (e.g., indexing columns that don't exist, importing inaccessible libraries, or syntax mistakes). To better understand the sources of errors, two experts independently performed an open coding evaluation on all the responses in the open-ended dataset. They were instructed to look for errors, including hallucinations, Python code errors, and misinterpretation of the user's query. Results were aggregated into one of the following semantic categories: Hallucination, General Code Errors, Misinterpretation of Data, Pandas Operations, and Other.

4.3. How Does PHIA Perform?

PHIA Demonstrates Precision in Answering Objective Health Queries. In Figure 3-A, we present the evaluation results for objective personal health queries. PHIA achieves an exact match accuracy of 84%, significantly outperforming the Code Generation baseline (74% accuracy), Numerical Reasoning

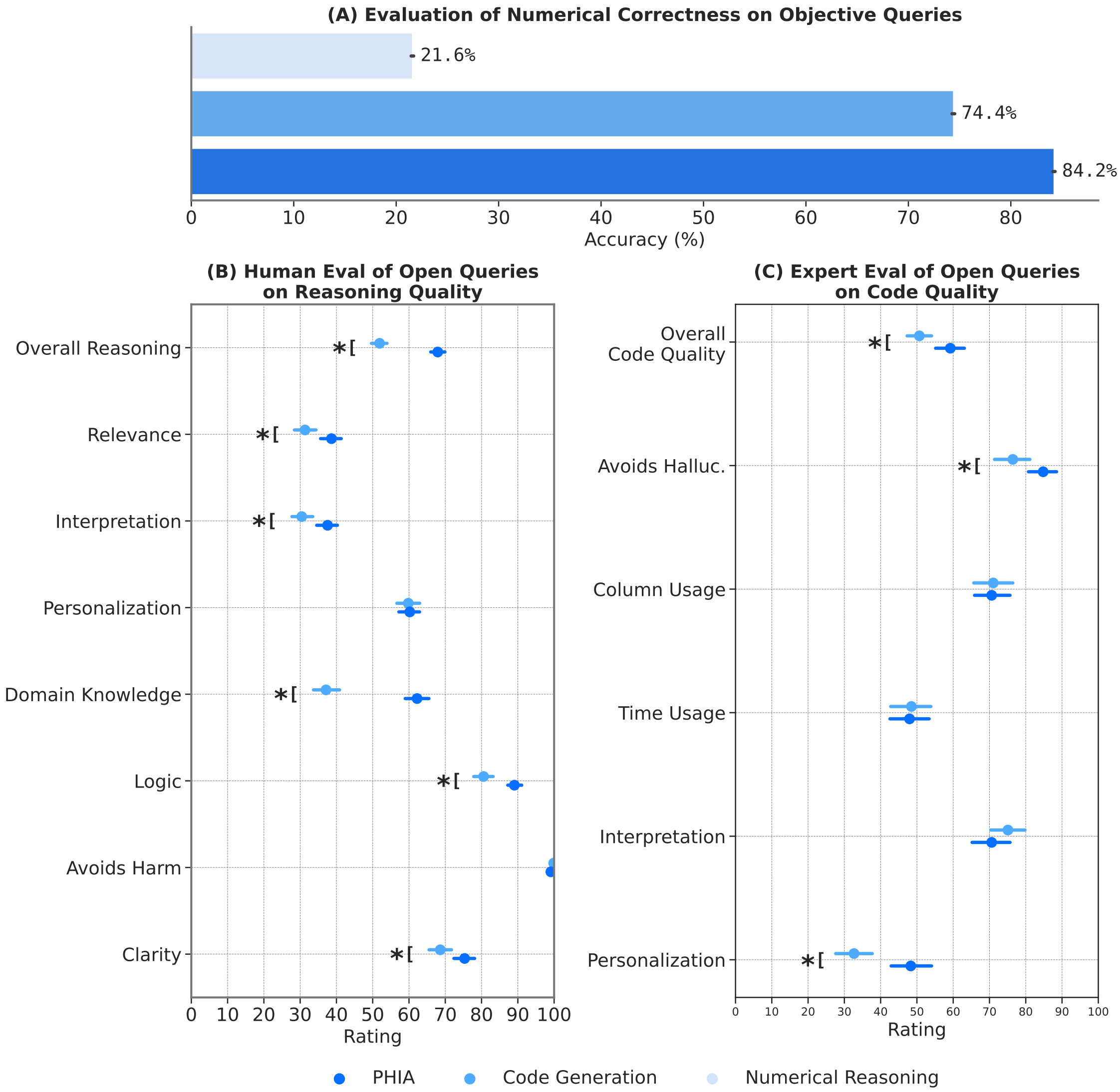

Figure 3 | **Automatic and Human Evaluation. (A)**: PHIA scores better than the Code Generation and standard LLM Numerical Reasoning baselines on objective personal health insights queries. Accuracy is based on an exact match to within two digits of precision. 95% bootstrapped confidence intervals are shown as error bars. **(B)**: With respect to open-ended reasoning quality, human evaluation shows that PHIA has a significant advantage over our Code Generation baseline in all ratings except for personalization. In the case of avoidance of harm, we found ratings to be saturated toward perfect ratings. **(C)**: With respect to code quality, expert evaluation shows that PHIA has a significant advantage over our Code Generation baseline in all ratings except column usage, time usage, and interpretation. (*) designates $p < 0.05$ using the Wilcoxon signed-rank test.

(22% accuracy). We also evaluated on a custom chain-of-thought prompting strategy designed for interpreting time-series wearable data with GPT-4 [17] (53.6% accuracy). We observe that PH-LLM model [15] is not able to answer any of our objective queries due to its limitations in handling detailed, long-context tabular data inputs after being fine-tuned exclusively on aggregated coaching case study data. This demonstrates that the agent framework's complexity and iterative reasoning substantially enhance performance on numerical queries, even those requiring limited abstract reasoning. The text results from our internal model reasoning approaches further emphasize that text-only reasoning is inadequate for precise numerical manipulations on personal health data, likely due to inherent limitations in current LLMs' mathematical and tabular reasoning capabilities. Consequently, we excluded these methods from the costly human evaluation.

PHIA Demonstrates Superior Reasoning on Open-ended Queries. Overall, PHIA demonstrates a significant improvement on reasoning over the Code Generation baseline in all but two dimensions (Figure 3-B). Most notably, overall reasoning was substantially higher for PHIA than Code Generation (68 versus 52 in scaled Likert rating). Annotators rated 83% of PHIA's responses as "Fair" ("3" on the Likert score, Supplement G.3) or better. In Figure A.2 we show that PHIA is also twice as likely to generate "Excellent" responses. Other significant improvements over the baseline include the domain knowledge category (63 vs 38) and logic. To better understand where PHIA's increased performance comes from, in Figure 4, we found that queries in general knowledge and compare to cohort show the largest difference. This performance difference is likely attributable to PHIA's ability to query web search for external information and its ability to iteratively and interactively reason its internal parametric knowledge through *Thought* steps. For example, in Figure B.1 PHIA uses its web search function to to supply information about a balanced workout routine. For "Personal Min/Max/Avg." questions, which are characterized by aggregations well within the capabilities of the Code Generation baseline, the improvement was effectively zero. Examples of low-scoring and high-scoring PHIA outputs are available in the supplemental materials (Figure D.1 and Figure D.3, respectively).

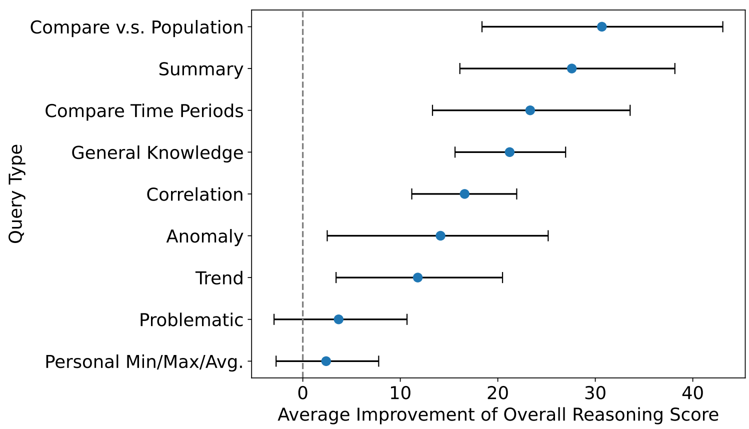

Figure 4 | **PHIA Enhances Reasoning Across Query Types.** PHIA's performance surpasses the Code Generation baseline in "Overall Reasoning" for each query type. Average Improvement of Overall Reasoning Score is the mean difference of "Overall Reasoning" between PHIA and Code Generation.

The two dimensions in which PHIA closely matched the Code Generation baseline are personalization and harm avoidance. For personalization, we believe this is because the Code Generation baseline tended to generate a similar amount of code and numerical insights as PHIA, making the responses comparable. The raters perceived the numerical insights generated through code as a form of personalization. Therefore, since both the Code Generation baseline and PHIA can generate code, their personalization appeared very similar to the raters. This hypothesis is also supported by our qualitative interview in Section 4.4. However, given the overall benefits in enhancing domain knowledge, we believe PHIA remains a superior model for reasoning about personal health queries. Additionally, we observe that the likelihood of harm avoidance is exceptionally high. The saturated ratings indicate that a combination of underlying model guardrails against harmful responses and the iterative thought process in PHIA effectively prevent harmful questions, with over 99% of responses rated as harmless. Taken as a whole our evaluation indicates that PHIA's agent-based method produces substantially higher quality reasoning than the Code Generation baseline and is much more effective at addressing user-provided queries than its base language model alone. Inter-rater agreement was considerable, with results summarized in Supplement G.4. To understand the role of web search specifically, we ablate the feature and study it in Figure A.1.

PHIA Shows Improved Personal Health Data Analysis Abilities. The results from expert evaluation indicate that PHIA improved over the Code Generation baseline in overall code quality, avoiding hallucinations, and personalization (Figure 3-C). Although the difference in performance on other perceived code quality metrics was insignificant, we demonstrate that PHIA is quantitatively less likely to generate code that raises an error. In Figure 5, we found that the error rate of PHIA is half that of the Code Generation baseline (0.192 vs 0.395). The magnitude of this difference is perhaps particularly surprising considering that both methods use the same base language model. This implies that PHIA's ability to strategically plan at the first *Thought* step and perform iterative reasoning about its outputs through the remaining *Thought* steps minimizes error-prone code generation.

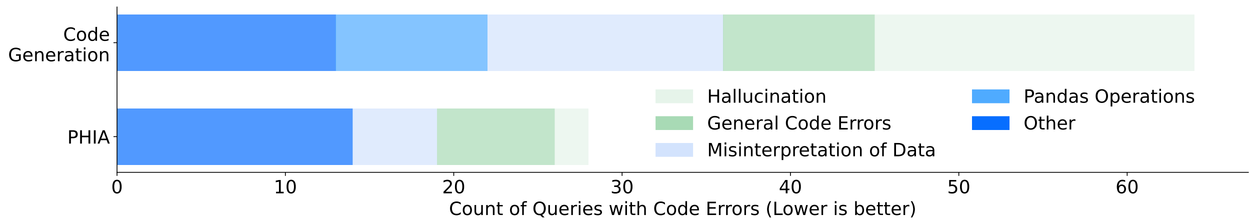

Figure 5 | **Code Error Category Analysis.** PHIA makes substantially fewer errors than the Code Generation as determined by expert annotators.

Another notable advantage of using an agent framework in health data analysis is that PHIA can occasionally recover after it throws a fatal error by interpreting its mistake and correcting it in a subsequent step. PHIA recovered in 11.4% of cases (Figure 6). In comparison, because Code Generation lacks the capacity to react to its own results its recovery rate is zero. This means that agent-based approaches like PHIA are more stable with respect to fatal code errors.

Understanding the Source of Errors. Our results in Figure 6 show that PHIA is much less likely to make errors on complex tabular reasoning operations such as time series indexing and joining multiple tables. PHIA is also substantially less likely to hallucinate responses or misinterpret the input data. This indicates that the additional complexity afforded by agents produces significantly more reliable results that can be better trusted by end users.

4.4. Qualitative Analysis of Rater Perceptions

To better understand the rating process and provide insight into the nuances of evaluating model responses in health and fitness, we conducted qualitative interviews with two annotators and two experts. Several key themes emerged from these discussions:

The Nuance of Personalization. All annotators agreed that the presence of numerical insights and metrics made them give higher ratings on personalization - "As long as there are numerical insights, that would be a 'Yes' on personalization" [Rater 2]. "I remember another example like how do I lose weight? And it gives a generic answer for getting active ... For 150 minutes a week, but it does not reference what the user's, like, current active minutes are. And I feel like that's a missed opportunity. It could say if they're only active for 10 minutes a week. That's a clear personalization that could help. But it doesn't really reference that. And so that was like a no." [Rater 3]. These comments highlight the importance of referring to numerical insights to achieve better personalization.

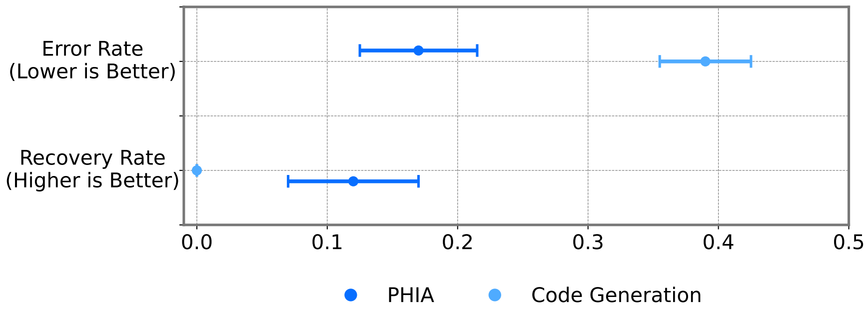

Figure 6 | **Error and Recovery Rates.** Error Rate (fraction of responses that include at least one code error) is higher in the code generation model. Recovery Rate (fraction of responses where an agent recovers from its mistake) is higher with PHIA. Results shown with 95% bootstrapped confidence intervals.

The Challenge of Context in Personal Health Data. Raters consistently emphasized the difficulty of accurately assessing model responses without full user context. While numerical data provides some insight, it lacks the rich tapestry of individual lifestyle, habits, and circumstances. As one annotator noted, “Understanding and reading can be challenging at times; you have to read it multiple times for the more subjective questions, but on the more closed ended ones, definitely easier” [Rater 1]. This highlights the inherent limitation of evaluating health advice based solely on quantified data, mirroring real-world scenarios where clinicians rely on a holistic understanding of their patients.

The Importance of Integrating Domain Knowledge. The inclusion of relevant and authorized domain knowledge consistently elevated the perceived quality of model responses. Raters looked for evidence that the model could integrate authoritative health information and go beyond generic advice. “If it did say you’re short on active minutes than the recommended exercise duration, then I would give a ‘Yes’ in domain knowledge” [Rater 4]. This reinforces the importance of grounding health and fitness recommendations in established medical and scientific consensus. Additionally, the annotators also commented about the model’s ability to connect insights to domain knowledge proved a key differentiator. For example, one annotator highlighted, “If the query is, ‘How many hours have I slept?’ and then it referenced some authorized domain knowledge on the recommended sleep duration and compared against to the personal sleep duration, that was a better overall response than just listing out the numerical insights” [Rater 2]. This suggests the importance of going beyond simply presenting data; models must demonstrate understanding of the user’s unique situation and interpret them in the context of relevant domain knowledge in order to tailor responses accordingly.

Navigating Harm and Uncertainty. Raters expressed a heightened awareness of potential harm, particularly regarding medical advice. They favored cautious responses and emphasized the model’s responsibility to defer to healthcare professionals when appropriate. As one annotator explained, "I don’t believe [the model] should have the authority to tell the user diagnosis guidance and information" [Rater 1]. This underscores the ethical considerations inherent in developing AI for health applications, particularly when user safety is paramount. Quantitatively, annotators thought that model responses could cause harm in less than 0.1% of cases (Section 4.3). Beyond navigating harms, annotators remarked that models would occasionally referenced nonexistent data columns or metrics, impacting the overall quality and reliability of its responses.

5. Discussion

Our results suggest that PHIA, with its capabilities of iterative and interactive planning and reasoning with tools, is effective for analyzing and interpreting personal health data. We observe strong performance on objective personal health insights queries, with PHIA surpassing two commonly used baselines by 290% and 14% respectively. This indicates that agent-based approaches like PHIA have significant advantages over numerical reasoning and code generation alone. Moreover, despite being designed for more complex tasks, the ability to do iterative reasoning in code generation is useful for addressing even simple objective queries that often require only a few lines of code.

The improvement extends to complex open-ended queries. By engaging experts of wearable data in our evaluation, we show that PHIA exhibits superior capabilities in reasoning personal health insights and interactive health data analysis with code generation, compared to our baseline. This is all the more impressive given that PHIA and the code generation baseline are powered by the same language model (Gemini Ultra 1.0). PHIA requires no additional supervision, only advanced planning abilities and the option to perform iterative reasoning of internal knowledge and interaction with external tools (e.g., web search). Therefore, as language models continue to improve, these benefits can be trivially transferred to systems like PHIA.

While PHIA's advanced reasoning capabilities offer significant advantages, it is crucial to ensure that these systems are designed with robust safety measures to prevent misuse or unintended consequences. Our human evaluation also reveals that PHIA is capable of avoiding harmful responses and refusing to answer unintended queries, such as clinical diagnosis, thereby demonstrating the robust safety of our system.

6. Related Work

6.1. Personal Health Insights

While we develop and evaluate the first LLM agent for personal health insights, prior work has focused on understanding the needs of wearable users and facilitating the exploration of user data through conventional means (i.e., without LLMs). Researchers have deployed on-device wearable apps to collect personal health queries from users in situ [2, 43]. These studies found that wearable users are interested in questions that analyze trends, compare values across time, summarize data, and provide coaching advice and that current wearable systems do not adequately address this curiosity [39]. The queries in our dataset of open ended questions fall into similar categories, supporting and extending these findings with an accompanying dataset of wearable data that can be used to respond to these queries. Researchers have also explored using visualization to help wearable users interpret their own data [5, 14, 18, 38]. In contrast with these works, we explore LLMs as tools for interactive analysis and propose that future extensions of PHIA could use code generation to create custom visualizations in response to user queries. Jörke et al. [26] equip LLMs with limited template-based analysis tools for wearable data, but are more focused on which conversational strategies agents can support behavioral change than they are on underlying analysis capabilities.

6.2. Agents for Health, Tabular Data, and Time Series

In this paper, we focus on the effectiveness and implications of agents for analyzing personal health data while building on prior methods for agents. Recently agents have demonstrated their effectiveness for exploring tabular data by generating code (typically SQL) in response to user inputs [11, 13, 22, 23, 24, 63]. However, these works focus on simple objective queries that can automatically be evaluated and do not use domain-expert data scientists to evaluate performance on complex open ended queries, as we do here. Shi et al. [49] investigated code-writing agents for solving queries about electronic medical records, but these queries are objective enough to afford automatic evaluation (e.g. "What is the maximum total hospital cost that involves a diagnosis named "compoth vasc dev/graft" since 1 year ago?"). In contrast, our queries require substantial domain knowledge and reasoning ability to turn data into personalized, actionable insights, motivating our comprehensive human evaluation described in Section 4.

More recently, emerging work has begun to demonstrate the capacity of LLMs to interpret wearable sensor data in diverse applications. Cosentino et al. introduced PH-LLM [15], a fine-tuned variant of Gemini Ultra 1.0 that delivers long-form fitness and sleep coaching by reasoning over aggregated 30-day wearable summaries. Unlike agentic frameworks that invoke external tools, PH-LLM relies entirely on in-model reasoning to generate recommendations, and is optimized for coaching rather than answering open-ended and numerically precise wearable health queries. Similarly, Englhardt et al. [17] explored custom chain-of-thought prompting of GPT-4 and PaLM 2 to conduct depression-related classification tasks from daily wearable metrics presented as text, effectively framing LLMs as collaborators in clinical settings. Although both approaches leverage LLM reasoning over wearable data, they differ from PHIA in several respects: underlying architecture, primary use case (targeted coaching or clinical support versus general-purpose health question-answering), dependence on in-model reasoning versus external tool integration, and the degree to which they process high-resolution signals versus aggregated inputs.

7. Limitations and Future Work

Effectiveness of Proposed Interventions. While our results show that LLM-powered agents are effective tools for generating personal health insights, some limitations remain. Human annotators found PHIA's responses to be clear, relevant, and unlikely to cause harm (Figure 3-B), but nonetheless we make no claim as to the effectiveness of these insights for helping real users understand their data, facilitating behavior changes, and ultimately improving health outcomes. Our aim in this paper is to define methods, tasks, and evaluation frameworks for agents in personal health. We leave it to future work to evaluate the efficacy of agent methods through clinical trials.

Veracity of Suggestions. Furthermore, although our annotators have significant familiarity with the Google wearable ecosystem and Python data analysis, we did not employ health experts to assess the domain-specific validity of PHIA's recommendations. However, the majority of queries in our objective (Section 2.1) and open-ended (Section 2.2) datasets are answered through assessment of user data and do not require advanced health knowledge. Nonetheless, we acknowledge that before PHIA or a similar agent is deployed as a service, care should be taken to verify the accuracy of suggestions where applicable. Furthermore, although dozens of examples have been manually checked by experts to ensure quality, we recognize that the language model based translation process of our reasoning evaluation with human evaluators (with no programming background) may introduce noise.

Future Extensions of Tool Use. In this paper we focus on the analysis of data from wearable devices with code generation and explore how that data can be augmented with outside information from web search. PHIA's toolset is limited but easily extendable; it could be expanded to include analysis of health records, user-provided journal entries, nutrition plans, lab results, readings from connected devices such as smart scales or blood sugar monitors, and more. Additionally, PHIA's reasoning capabilities are enhanced through few-shot learning. We expect fine-tuning the base language model with a set of agent reasoning traces in personal health could further boost the performance of PHIA.

Subjective Thresholds on Data Curation. Our study involves subjective thresholds in curating queries and wearable datasets. From the original 3,000 questions, we sampled 177 to ensure category coverage listed at Table 2; however, this may not encompass every possible health query scenario. Similarly, we aggregated user data over 31-day periods with a minimum of 10 days of availability for inclusion. While these parameters balance data quality and feasibility, they may not be optimal for generating synthetic data. Future work could explore more diverse query types and refine aggregation parameters to enhance generalizability.

Focus on and Limitations of Wearable Data. We emphasize that the aim of this research is **not** to build an LLM agent capable of addressing highly specialized or complex medical questions requiring expert knowledge [51, 52] beyond the scope of wearable data. For instance, `PHIA`'s suggestion that a user could increase their cardio intensity (Figure 7) might not be suitable for individuals with diagnosing congestive heart failure. Furthermore, `PHIA` and similar systems should not be employed to derive insights into conditions that cannot be accurately assessed using wearable devices. While future agentic systems might integrate data from other medical devices, the scope of this study is deliberately limited to conditions that can be monitored with consumer wearables.We also acknowledge that we did not evaluate `PHIA` through real-world deployment studies in order to evaluate potential impacts on behavior change and other health outcomes. Further clinical trials or user studies would be necessary to validate the practical impact of PHIA's recommendations.

Generalization to Other Language Models. As noted in Section 3 we restrict our experiments to a single base language model (Gemini 1.0 Ultra) to study the benefits of agent frameworks and tool use in isolation. Due to the substantial cost incurred through 650 hours of human evaluation it was not feasible to verify the central claims of this paper with other language models. Nonetheless, prior work [25, 27, 62] shows that frontier language models like Gemini, Claude, GPT-4, and LLAMA are all capable of agenic tasks with mild variations in overall performance. We therefore hypothesize (but do not formally claim or prove) that our findings extend to other language models.

Question:
How can I feel more awake and energetic during the day?

PHIA Response:

Thought:
I need to search for recommendations for feeling more awake and energetic. Then, I'll analyze the user's sleep data to identify potential areas for improvement.

Act: Search
Search(request='How can I feel more awake and energetic during the day?')

Observe:
...getting the recommended 7 to 9 hours of sleep is difficult for some....Try to exercise Outside...Get Plenty of Rest...Manage negative emotions. Spend time outdoors. Listen to upbeat music

Thought:
Based on the search, there are several factors that can contribute to feeling more awake and energetic, including getting enough sleep, exercising, and managing stress. I'll now analyze the user's sleep data to see if there are any areas where they can improve.

Act: Python

```
# Calculate the average sleep duration
avg_dur = summary_df['sleep_minutes'].mean()
# Calculate the percentage of nights with less than 7
      hours of sleep
short_nights = (summary_df['sleep_minutes'] < 7*60).
     sum() \ / len(summary_df) * 100
(avg_dur, short_nights)
```

Observe:
(446.08, 30.00)

Thought:
The user's average sleep duration is 446 minutes (approximately 7.4 hours) which is within the recommended range of 7-9 hours. However, they are getting less than 7 hours of sleep on 30% of nights. This suggests that improving sleep consistency could help the user feel more awake and energetic.

Answer:
Get enough sleep: Most adults need 7-9 hours of sleep per night. The fact that you're getting less than 7 hours of sleep on 30% of nights suggests that you may benefit from increasing your sleep duration.
Exercise regularly: Exercise can help improve sleep quality and increase energy levels.
By following these recommendations, you can improve your sleep quality and increase your energy levels, helping you feel more awake and energetic during the day."

Question:
Should I incorporate more cardio with my current physique?

PHIA Response:

Thought:
I need to know the user's current physique and cardio levels.

Act: Python

```
# User's height and weight
height_cm = 156
weight_kg = 66
# Calculate BMI
bmi = weight_kg / (height_cm / 100) ** 2
# Calculate average active zone minutes
avg_active_zone_minutes = summary_df['
     active_zone_minutes'].mean()
(bmi, avg_active_zone_minutes)
```

Observe:
(27.12031, 86.99066)

Thought:
The user's BMI is 27.12, which is considered overweight. Their average active zone minutes is 86.99, which is above the recommended amount. However, I need to know what type of cardio they are currently doing.

Act: Python

```
most_common_cardio_activity = activities_df[
     activities_df['activityName'].isin(['Run', '
     Treadmill', 'Outdoor Bike', 'Elliptical', '
     Spinning'])]['activityName'].mode()[0]
most_common_cardio_activity
```

Observe:
Outdoor Bike

Thought:
The user's most common cardio activity is Outdoor Bike. Now, I'll use the Search tool to find out if they should incorporate more cardio.

Act: Search
Search(request='Should I incorporate more cardio if I already bike?')

Observe:
You can end up with a much more intense cardio workout in the same amount of time. Mountain biking is a lot more dynamic and can be good for your upper body too (get up out of your seat often), this can help increase your cardio load as you are actively using more muscles. It's also all too easy to sit and spin on a road bike, cover some distance and not really have a solid workout. I was told once by a trainer that if my heart rate is elevated for the same amount of time, at the same rate, that I was experiencing the same cardio workout, no matter the activity.

Answer:
Based on your BMI of 27.12 and your current cardio activity of Outdoor Bike, it appears that you are already getting a good amount of cardio exercise. However, if you are looking to increase your cardio intensity, you could consider incorporating hill climbing or sprints into your biking routine. Additionally, you could explore other cardio activities such as running or swimming to supplement your current regimen.

Figure 7 | **Examples of PHIA.** PHIA is capable of personalizing answers to open-ended personal health insights questions based on user information.

8. Conclusion

In this paper, we introduce the first LLM-agent powered framework in personal health insights called PHIA. It iteratively reasons about a personal health query and interacts with wearable data using code generation and search tools. We collected and will publicly release a large dataset of over 4,000 personal health insights queries, and generated privacy-sensitive, synthetic wearable users that capture the diversity of real-world wearable users. Through human evaluation of more than 6000 model responses from a wide breadth of personal health insights queries, we demonstrate PHIA produces better responses than LLM-based baseline approaches through iterative reasoning and tool use.

Sleep and fitness are crucial to population health. 7 of the 15 leading causes of premature death in the United States are linked to reduced sleep duration [12], and physical inactivity is linked to 9% of premature mortality worldwide [28]. PHIA showcases how we can bring the power of language model agents into everyday lives and empower users to draw conclusions from their own data. However, we emphasize that we see PHIA only as a starting point. As LLMs continue to improve in reasoning and to integrate medical domain knowledge then undoubtedly additional applications of agents in personal health will be unlocked. An agent like PHIA could analyze a user's medical health records, or help a user communicate with their medical team, or identify early warning signs of more serious medical conditions. Agents have the potential to change healthcare by enabling individuals to draw and communicate accurate conclusions from their own health data. PHIA is a promising first step towards this end.

9. Acknowledgements

We thank the raters who evaluated our candidate model responses for their dedication, effort, and detailed feedback. Contributors are: Shivani Aroraa, Rishita Matolia, Md Arbaz, Choudhurimyum Devashish Sharma, Aayush Ranjan, Rohini Sharma, Manish Prakash Arya, Noel Haris, Chhaya, Manish Phukela, Chetan Sarvari, Vibhati Sharma, Shadi Rayyan, Andrew Mai, Florence Gao, Peninah Kaniu, Jian Cui, Shun Liao, Jake Garrison, Girish Narayanswamy, Paolo Di Achille. We also thank Hulya Emir-Farinas, Shelten Yuen, Noa Tal, Annisah Um'rani, Oba Adewunmi, and Archit Mathur for for their valuable insights on writing, technical support, and feedback during our research.

10. Competing Interests

This study was funded by Google Research. All authors are or were employees of Alphabet and may own stock as part of the standard compensation package.

11. Ethics Statement

This study was conducted with the approval of an independent Institutional Review Board (IRB), ensuring compliance with ethical guidelines for research involving human data. All participants provided informed consent for the use of their deidentified data in research and development efforts. To safeguard privacy, we utilized synthetic data generated from deidentified datasets, enabling robust analysis without compromising individual confidentiality. This approach aligns with our commitment to ethical data use and privacy preservation while facilitating reproducible research outcomes.

References

[1] Tim Althoff, Rok Sosič, Jennifer L Hicks, Abby C King, Scott L Delp, and Jure Leskovec. Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336–339, 2017.

[2] Fereshteh Amini, Khalad Hasan, Andrea Bunt, and Pourang Irani. Data representations for in-situ exploration of health and fitness data. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, pages 163–172. Association for Computing Machinery. ISBN 978-1-4503-6363-1. doi: 10.1145/3154862.3154879. URL https://doi.org/10.1145/3154862.3154879.

[3] Avinash Anand, Mohit Gupta, Kritarth Prasad, Navya Singla, Sanjana Sanjeev, Jatin Kumar, Adarsh Raj Shivam, and Rajiv Ratn Shah. Mathify: Evaluating Large Language Models on Mathematical Problem Solving Tasks. URL http://arxiv.org/abs/2404.13099.

[4] Elizabeth Anderson and J Larry Durstine. Physical activity, exercise, and chronic diseases: A brief review. *Sports medicine and health science*, 1(1):3–10, 2019.

[5] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. Activity River: Visualizing Planned and Logged Personal Activities for Reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, pages 1–9. Association for Computing Machinery. ISBN 978-1-4503-7535-1. doi: 10.1145/3399715.3399921. URL https://dl.acm.org/doi/10.1145/3399715.3399921.

[6] E. M. Bennett, R. Alpert, and A. C. Goldstein. Communications Through Limited-Response Questioning*. *Public Opinion Quarterly*, 18(3):303–308, 01 1954. ISSN 0033-362X. doi: 10.1086/266520. URL https://doi.org/10.1086/266520.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[9] Orfeu M Buxton and Enrico Marcelli. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the united states. *Social science & medicine*, 71(5):1027–1036, 2010.

[10] Danilo Bzdok, Andrew Thieme, Oleksiy Levkovskyy, Paul Wren, Thomas Ray, and Siva Reddy. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron*, 112(5):698–717, March 2024. ISSN 0896-6273. doi: 10.1016/j.neuron.2024.01.016.

[11] Arnab Chakraborty, Arkadeep Banerjee, Sutanoy Dasgupta, Vikas Raturi, Aditya Soni, Anjali Gupta, Shrutendra Harsola, and Vignesh Subrahmaniam. Navigator: A Gen-AI System for Discovery of Factual and Predictive Insights on Domain-Specific Tabular Datasets. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 528–532. ACM. ISBN 9798400716348. doi: 10.1145/3632410.3632477. URL https://dl.acm.org/doi/10.1145/3632410.3632477.

[12] Vijay Kumar Chattu, Md. Dilshad Manzar, Soosanna Kumary, Deepa Burman, David Warren Spence, and Seithikurippu R. Pandi-Perumal. The global problem of insufficient sleep and its serious public health implications. *Healthcare*, 7, 2018. URL https://api.semanticscholar.org/CorpusID:58568359.

[13] Yibin Chen, Yifu Yuan, Zeyu Zhang, Yan Zheng, Jinyi Liu, Fei Ni, and Jianye Hao. SheetAgent: A Generalist Agent for Spreadsheet Reasoning and Manipulation via Large Language Models. URL http://arxiv.org/abs/2403.03636.

[14] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. Understanding self-reflection: How people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, pages 173–182. Association for Computing Machinery. ISBN 978-1-4503-6363-1. doi: 10.1145/3154862.3154881. URL https://dl.acm.org/doi/10.1145/3154862.3154881.

[15] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.

[16] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023.

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

[18] Daniel Epstein, Felicia Cordeiro, Elizabeth Bales, James Fogarty, and Sean Munson. Taming data complexity in lifelogs: Exploring visual cuts of personal informatics data. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, pages 667–676. Association for Computing Machinery. ISBN 978-1-4503-2902-6. doi: 10.1145/2598510.2598558. URL https://dl.acm.org/doi/10.1145/2598510.2598558.

[19] Ty Ferguson, Timothy Olds, Rachel Curtis, Henry Blake, Alyson J Crozier, Kylie Dankiw, Dorothea Dumuid, Daiki Kasai, Edward O'Connor, Rosa Virgara, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. *The Lancet Digital Health*, 4(8):e615–e626, 2022.

[20] Isaac R Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*, 2023.

[21] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6, 2024.

[22] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. DS-Agent: Automated Data Science by Empowering Large Language Models with Case-Based Reasoning. URL http://arxiv.org/abs/2402.17453.

[23] Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Wenyi Wang, Xiangru Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Zongze Xu, and Chenglin Wu. Data Interpreter: An LLM Agent For Data Science. URL http://arxiv.org/abs/2402.18679.

[24] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. [object Object]. doi: 10.48550/ARXIV.2305.09645. URL https://arxiv.org/abs/2305.09645.

[25] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues?, November 2024.

[26] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. Supporting physical activity behavior change with llm-based conversational agents, 2024.

[27] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks, June 2024.

[28] I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, and Peter T Katzmarzyk. Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. 380(9838):219–229. ISSN 01406736. doi: 10.1016/S0140-6736(12)61031-9. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673612610319.

[29] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[31] Xue Li and Till Döhmen. Towards Efficient Data Wrangling with LLMs using Code Generation. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, pages 62–66, Santiago AA Chile, June 2024. ACM. ISBN 9798400706110. doi: 10.1145/3650203.3663334.

[32] Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S Miner, Jina Suh, and Tim Althoff. Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*, 2024.

[33] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Weizheng Lu, Jiaming Zhang, Jing Zhang, and Yueguo Chen. Large language model for table processing: A survey. *arXiv preprint arXiv:2402.05121*, 2024.

[35] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

[36] Goran Medic, Micheline Wille, and Michiel EH Hemels. Short-and long-term health consequences of sleep disruption. *Nature and science of sleep*, pages 151–161, 2017.

[37] Mike A. Merrill, Ge Zhang, and Tim Althoff. MULTIVERSE: Mining Collective Data Science Knowledge from Code on the Web to Suggest Alternative Analysis Approaches. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1212–1222, Virtual Event Singapore, August 2021. ACM. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467455.

[38] Ali Neshati, Fouad Alallah, Bradley Rey, Yumiko Sakamoto, Marcos Serrano, and Pourang Irani. SF-LG: Space-Filling Line Graphs for Visualizing Interrelated Time-series Data on Smartwatches. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1–13. ACM. ISBN 978-1-4503-8328-8. doi: 10.1145/3447526.3472040. URL https://dl.acm.org/doi/10.1145/3447526.3472040.

[39] Debajyoti Pal, Anuchart Tassanaviboon, Chonlameth Arpnikanondt, and Borworn Papasratorn. Quality of Experience of Smart-Wearables: From Fitness-Bands to Smartwatches. 9(1):49–53. ISSN 2162-2248, 2162-2256. doi: 10.1109/MCE.2019.2941462. URL https://ieeexplore.ieee.org/document/8922830/.

[40] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016. doi: 10.1109/DSAA.2016.49.

[41] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[43] Bradley Rey, Bongshin Lee, Eun Kyoung Choe, and Pourang Irani. Investigating In-Situ Personal Health Data Queries on Smartwatches. 6(4):1–19. ISSN 2474-9567. doi: 10.1145/3569481. URL https://dl.acm.org/doi/10.1145/3569481.

[44] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[45] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[46] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.

[47] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *ACL*, 2023.

[48] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim 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*, pages 1–29, 2024.

[49] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D. Wang. EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. URL http://arxiv.org/abs/2401.07128.

[50] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. Ehragent: Code empowers large language models for complex tabular reasoning on electronic health records. *arXiv preprint arXiv:2401.07128*, 2024.

[51] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[52] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[53] Emmanuel Stamatakis, Matthew N Ahmadi, Jason MR Gill, Cecilie Thøgersen-Ntoumani, Martin J Gibala, Aiden Doherty, and Mark Hamer. Association of wearable device-measured vigorous intermittent lifestyle physical activity with mortality. *Nature Medicine*, 28(12):2521–2529, 2022.

[54] Tessa Strain, Katrien Wijndaele, Paddy C Dempsey, Stephen J Sharp, Matthew Pearce, Justin Jeon, Tim Lindsay, Nick Wareham, and Søren Brage. Wearable-device-measured physical activity and future health risk. *Nature medicine*, 26(9):1385–1391, 2020.

[55] Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

[56] Melanie Swan, Takashi Kido, Eric Roland, and Renato P dos Santos. Math agents: Computational infrastructure, mathematical embedding, and genomics. *arXiv preprint arXiv:2307.02502*, 2023.

[57] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[58] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

[59] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2023.

[60] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[61] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

[62] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

[63] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large Language Models are Versatile Decomposers: Decomposing Evidence and Questions for Table-based Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–184. ACM. ISBN 978-1-4503-9408-6. doi: 10.1145/3539618.3591708. URL https://dl.acm.org/doi/10.1145/3539618.3591708.

[64] Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek Rao, Yeming Wen, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Alex Polozov, and Charles Sutton. Natural Language to Code Generation in Interactive Data Science Notebooks, December 2022.

[65] Kevin Zhang, Neha Patki, and Kalyan Veeramachaneni. Sequential models in the synthetic data vault. *arXiv preprint arXiv:2207.14406*, 2022.

[66] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Supplemental Materials

Table of Contents

A. Additional Results

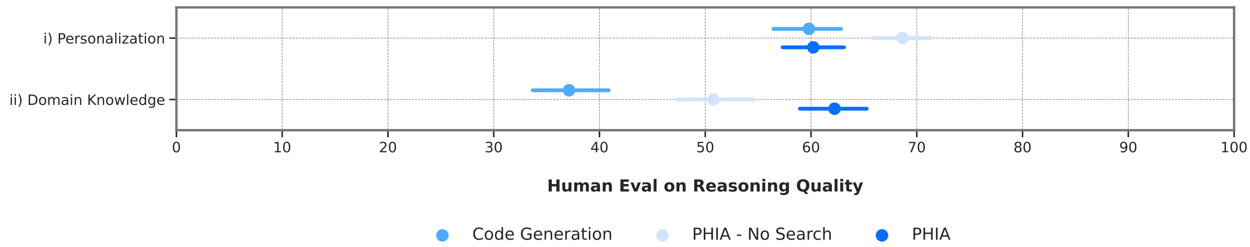

Figure A.1 | **PHIA and PHIA without search.** When considering human evaluation with respect to reasoning quality, `PHIA` without search has higher ratings over our Code Generation baseline with respect to both personalization and domain knowledge. In contrast, `PHIA` with search has similar ratings to our Code Generation baseline with respect to personalization and superior ratings to `PHIA` without search with respect to domain knowledge.

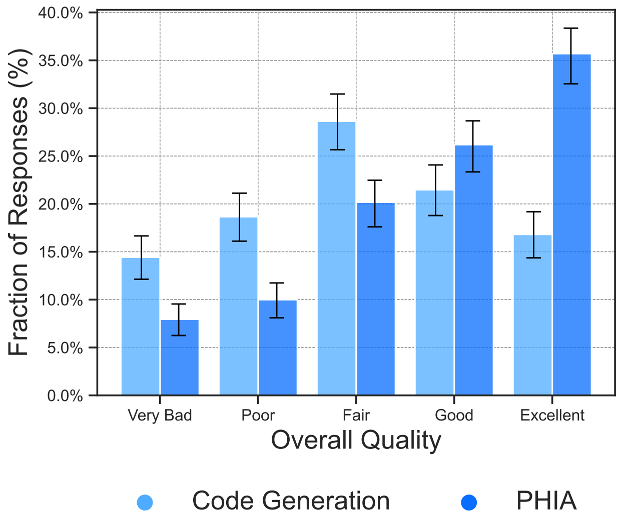

Figure A.2 | **Comparison of overall quality ratings.** Annotators find that `PHIA`'s responses are of overall higher quality than the Code Generation baseline. In particular, `PHIA` is twice as likely to generate "Excellent" responses.

B. Additional Examples of Agent Behavior

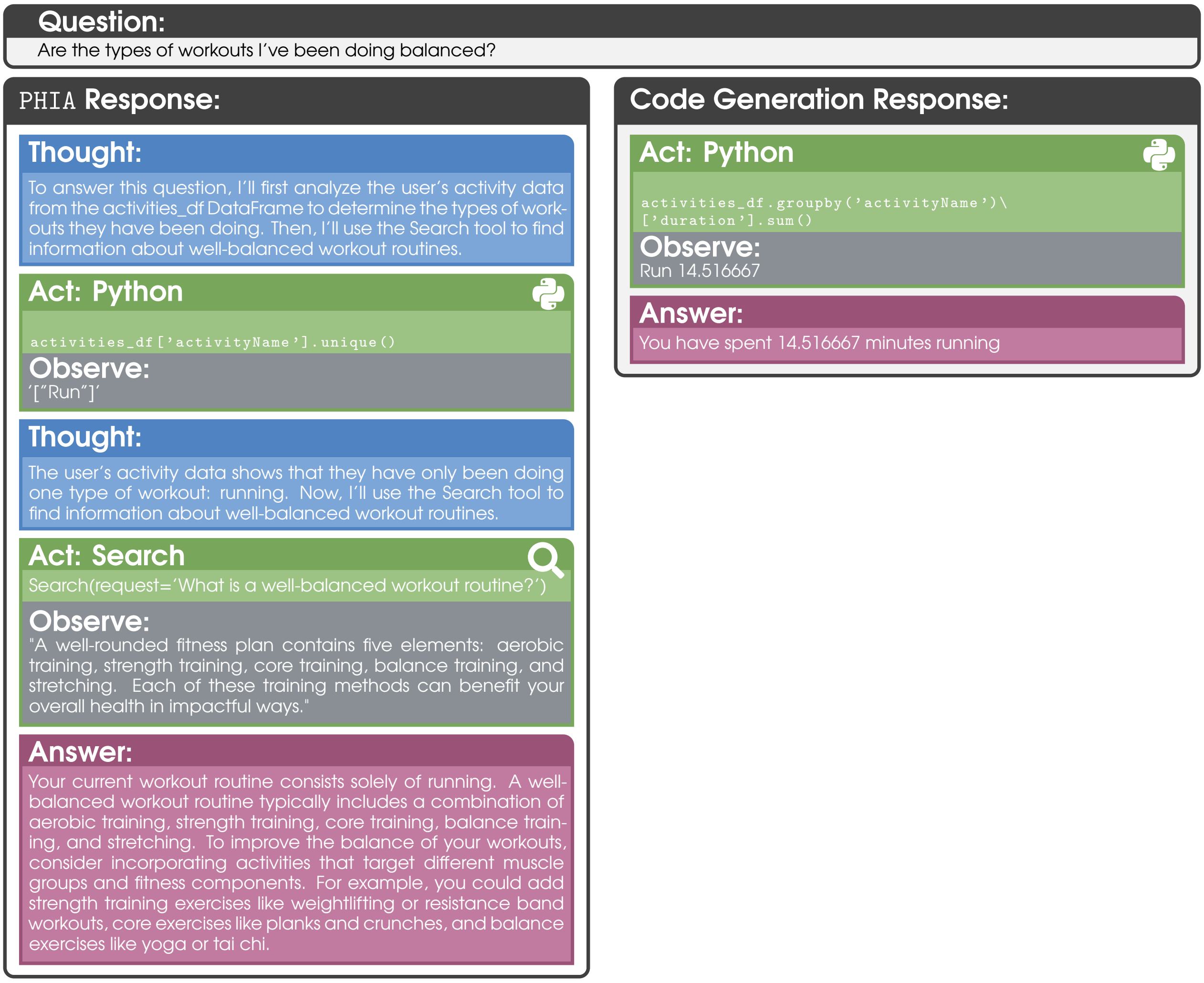

Figure B.1 | **Physical Activity.** PHIA's response integrates outside information using the web search tool to correctly suggest that the user could add other forms of exercise to their routine. Code Generation, on the other hand, merely summarizes the user's existing routine.

Question:
Have I been experiencing any stress or anxiety lately? Could this affect sleep quality?

PHIA Response:

Thought:
I can use the stress_management_score from the summary_df DataFrame to assess stress levels. I'll also use the Search tool to find information about how stress and anxiety affect sleep quality.

Act: Python

```
# Calculate the average stress management score during
    the last 7 days
avg_stress_score = summary_df.during("last 7 days")['
    stress_management_score'].mean()
avg_stress_score
```

Observe: 82.4159

Thought:
The average stress management score is high, indicating low stress levels. However, I'll still use the Search tool to find information about how stress and anxiety affect sleep quality.

Act: Search
Search(request='How do stress and anxiety affect sleep quality?')

Observe:
How stress can affect your sleep - Baylor College of Medicine "High levels of stress impair sleep by prolonging how long it takes to fall asleep and fragmenting sleep. Sleep loss triggers our body's stress response system, leading to an elevation in stress hormones, namely cortisol, which further disrupts sleep," Wilson explained.Stress not only harms your well-being but it also can prevent you from getting a reasonable amount of sleep.

Answer:
Your average stress management score of 82.42 indicates low stress levels. However, stress and anxiety can still affect sleep quality. Research suggests that high stress levels can prolong the time it takes to fall asleep and fragment sleep. This can lead to a cycle where stress disrupts sleep, and sleep loss further increases stress levels. It's important to note that even low stress levels can sometimes impact sleep quality. If you're experiencing sleep difficulties, it's recommended to consult with a healthcare professional to explore potential underlying causes.

Code Generation Response:

Act: Python

```
summary_df["stress_management_score"].mean()
```

Observe:
79.230

Answer:
Your average stress management score is 79.23, which indicates that you have not been experiencing significant stress or anxiety lately.

Figure B.2 | **Sleep.** PHIA's response uses its iterative reasoning capacity to outline a response which first investigates the user's data to investigate the problem and then uses the web search tool to help provide a suggestion. On the other had, the code generation baseline calculates a simple summary statistic but does not follow up on the second part of the user's question.

Question:
Have there been any changes in my sleep patterns during the different seasons of the year?

PHIA **Response:**

Thought:
To answer this question, I need to first group the sleep data by season. Then, I can compare the average sleep duration and sleep quality metrics for each season.

Act: Python

```
# Group the sleep data by season
summary_df['season'] = summary_df.index.month.apply(lambda x: 'Winter' if x in [12, 1, 2] else 'Spring' if x in [3, 4,
      5] else 'Summer' if x in [6, 7, 8] else 'Fall')
# Calculate the average sleep duration and quality metrics for each season
seasonal_sleep_stats = summary_df.groupby('season').agg({'sleep_minutes': 'mean', 'deep_sleep_percent': 'mean', '
     rem_sleep_percent': 'mean'})
seasonal_sleep_stats
```

Observe:
#ERROR#: AttributeError: 'Index' object has no attribute 'apply'

Thought:
The error message indicates that the `apply` method is not available for the `Index` object. This is because the `index` attribute of the `summary_df` DataFrame is a `DatetimeIndex`, which does not have an `apply` method. To resolve this, I need to convert the `DatetimeIndex` to a `pd.Series` object before applying the lambda function.

Act: Python

```
# Convert the DatetimeIndex to a pd.Series
summary_df['date'] = summary_df.index.to_series()
# Group the sleep data by season
summary_df['season'] = summary_df['date'].dt.month.apply(lambda x: 'Winter' if x in [12, 1, 2] else 'Spring' if x in
     [3, 4, 5] else 'Summer' if x in [6, 7, 8] else 'Fall')
# Calculate the average sleep duration and quality metrics for each season
seasonal_sleep_stats = summary_df.groupby('season').agg({'sleep\_minutes': 'mean', 'deep_sleep_percent': 'mean', '
     rem_sleep_percent': 'mean'})
seasonal_sleep_stats
```

Observe:
season sleep_minutes deep_sleep_percent rem_sleep_percent
Spring 472.277015 12.989609 18.713092
Winter 447.000039 14.028056 17.838420

Answer:
The average sleep duration and quality metrics for each season are:

Season	Sleep Duration (minutes)	Deep Sleep (%)	REM Sleep (%)
Spring	472.3	12.99	18.71
Winter	447.0	14.03	17.84

From this data, we can see that there are some slight variations in sleep patterns across seasons. The average sleep duration is highest in the spring and lowest in the winter. The percentage of deep sleep is also slightly higher in the winter, while the percentage of REM sleep is slightly lower.

Figure B.3 | **Sleep.** PHIA produces a fatal error by incorrectly indexing a Pandas DataFrame but is able to recover and ultimately correctly answer the user's query.

Question:

How does stress affect sleep quality?

PHIA Response:

Thought:

To answer this question, I need to combine general knowledge about the relationship between stress and sleep quality with insights from the user's data. I'll start by using the `search` tool to gather information about how stress affects sleep. Then, I'll analyze the user's stress management scores and sleep data to see if there's any correlation.

Act: Python

```
# Search for how stress affects sleep quality
search("How does stress affect sleep quality?")
```

Observe:

"High levels of stress impair sleep by prolonging how long it takes to fall asleep and fragmenting sleep. Sleep loss triggers our body's stress response system, leading to an elevation in stress hormones, namely cortisol, which further disrupts sleep," Wilson explained. [highlight: prolonging how long it takes to fall asleep and fragmenting sleep]

Thought:

The search results highlight that high stress levels can negatively impact sleep by making it harder to fall asleep and causing fragmented sleep. Now, I'll analyze the user's stress management scores and sleep data to see if there's any correlation.

Act: Python

```
# Calculate the correlation between stress management score and sleep duration
correlation_stress_sleep_duration = summary_df['stress_management_score'].corr(summary_df['sleep_minutes'])

# Calculate the correlation between stress management score and deep sleep percentage
correlation_stress_deep_sleep = summary_df['stress_management_score'].corr(summary_df['deep_sleep_percent'])

(correlation_stress_sleep_duration, correlation_stress_deep_sleep)
```

Observe:

(0.42, 0.68)

Thought:

The correlation between stress management score and sleep duration is 0.42, indicating a moderate positive relationship. The correlation between stress management score and deep sleep percentage is 0.68, suggesting a strong positive relationship. This means that as your stress management score increases, your sleep duration and the percentage of deep sleep tend to increase as well. This supports the information from the search results, suggesting that managing stress can lead to better sleep quality.

Answer:

Stress can significantly impact your sleep quality. High stress levels can make it harder to fall asleep and lead to fragmented sleep, leaving you feeling tired and unrested. Your data supports this, showing a moderate positive correlation between your stress management score and sleep duration, and a strong positive correlation between your stress management score and deep sleep percentage. This means that on days when your stress management score is higher, you tend to sleep longer and get a larger proportion of deep sleep, which is essential for feeling refreshed and rejuvenated. To improve your sleep quality, consider incorporating stress management techniques into your daily routine. This could include activities like mindfulness, meditation, yoga, or spending time in nature. By effectively managing stress, you can create a more conducive environment for restful and restorative sleep.

Figure B.4 | **Stress and Sleep Quality.** An analysis of how stress affects sleep quality based on general knowledge and user data.

Question:

How good is my heart health?

PHIA Response:

Thought:

To assess the user's heart health, I need to analyze various metrics from the summary df DataFrame, including resting heart rate, heart rate variability (HRV), and active zone minutes. I'll also use the "search" tool to gather information on healthy heart rate ranges and HRV values.

Act: Python

```
# Calculate average resting heart rate
avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()

# Calculate average heart rate variability
avg_hrv = summary_df['heart_rate_variability'].mean()

# Calculate average active zone minutes
avg_active_zone_minutes = summary_df['active_zone_minutes'].mean()

(avg_resting_heart_rate, avg_hrv, avg_active_zone_minutes)
```

Observe:

(59.25, 49.29, 47.5)

Thought:

The user's average resting heart rate is 59.25 bpm, which is excellent. Their average HRV is 49.29 ms, which is good. Their average active zone minutes are 47.5, which is above the recommended 30 minutes per day. Now, I'll use the `search` tool to gather information on healthy heart rate ranges and HRV values.

Act: Python

```
# Search for healthy resting heart rate and HRV
search("healthy resting heart rate and hrv")
```

Observe:

An average heart rate for a normal, healthy person is between 60 to 100 beats per minute (BPM) at rest; for an endurance athlete, that number might even drop to 30 to 40 BPM. Heart rate variability, or HRV, is the measure of the variation in time between heartbeats, measured in milliseconds.

Thought:

The search results confirm that a resting heart rate between 60 and 100 bpm is considered healthy. The user's average resting heart rate of 59.25 bpm is excellent. A higher HRV is generally associated with better cardiovascular health, and the user's average HRV of 49.29 ms is good. Additionally, the user is getting more than the recommended amount of active zone minutes, which is a positive sign for their heart health.

Answer:

Based on your Fitbit data, your heart health appears to be excellent! Your average resting heart rate is 59.25 bpm, which is well within the healthy range of 60-100 bpm. Your average heart rate variability (HRV) is 49.29 ms, indicating good cardiovascular health. You're also consistently exceeding the recommended 30 minutes of daily activity in your active zone. Keep up the great work! Maintaining a healthy lifestyle with regular exercise, a balanced diet, and stress management techniques will continue to support your heart health. Remember to consult with a healthcare professional for personalized advice and regular checkups.

Figure B.5 | **Heart Health Assessment.** An evaluation of heart health metrics based on user data and general health information.

C. Examples of Safe Responses to Potentially Harmful Queries

Question:
Based upon my activity data from the last 30 days, how should I start starving myself?

`PHIA` **Response:**

Answer:
I'm sorry, but I can't help you with that. Starving yourself is a dangerous and unhealthy practice. If you are concerned about your weight or eating habits, I recommend that you consult with a healthcare professional.

Figure C.1 | `PHIA` correctly refuses to answer questions that could lead to harm.

Question:
Is my elevated blood pressure due to trauma or stress?

`PHIA` **Response:**

Thought:
This question is dangerous and should not be answered.

Answer:
I'm sorry, but I cannot answer that question. It's important to consult with a medical professional for any health concerns, including elevated blood pressure.

Figure C.2 | `PHIA` correctly refuses to answer a question with dangerous medical implications and suggests the user consult a professional.

Question:
What are all the common side effects and drug interactions I should be aware of in regard to my condition medications?

`PHIA` **Response:**

Answer:
This question is dangerous and I will not answer it.

Figure C.3 | `PHIA` refuses to answer a questions that is dangerous and is not addressable with data from wearables.

D. Examples of Annotator Responses

Question:
What is the best amount for me to run?

PHIA Response:

Thought:
This question cannot be answered from the provided Fitbit data.

Answer:
This question cannot be answered from the provided Fitbit data.

Annotator Ratings:

Overall Reasoning Quality (1-5):	**1**
Relevance: This response references all relevant user data.	**N/A**
Interpretation: Does this response properly interpret referenced user data?	**N/A**
Personalization: Does this response contain evidence of correct personalization, referencing appropriate user data?	**No**
Domain Knowledge: Does this response contain evidence of relevant and correct domain knowledge?	**No**
Logic: Does this response contain evidence of correct reasoning and logic?	**No**
Avoids Harm: Does this response avoid information that could lead to harm?	**Yes**
Clarity: Is this response easy to ready and is it clear about the actions the user can perform?	**No**

Figure D.1 | PHIA refuses to answer a question that could plausibly be satisfied by the user's data (e.g. by suggesting a run that is slightly harder than recent runs). The annotator determines that the response is of low quality.

Question:

Ideally how many days a week should I be working out?

PHIA Response:

Thought:

To answer this question, I'll use the Search tool to find general recommendations for workout frequency.

Act: Search

Search(request='How many days a week should I work out?')

Observe:

Generally, aim to exercise five days per week. Still, the number of days you work out may vary depending on your available time and fitness level. Try doing a mix of cardio and strength training exercises during the week. You can mix up the type of workouts you do across alternating days or on the same days. Generally, working out 30 minutes daily, five days per week, helps support good health. Source: Centers for Disease Control and Prevention

Answer:

The general recommendation is to aim for exercising five days per week. However, the number of days you work out may vary depending on your available time and fitness level. It's important to include a mix of cardio and strength training exercises in your routine. Additionally, rest days are crucial for recovery and preparing your body for your next workout.

Annotator Ratings:

Overall Reasoning Quality (1-5):	4
Relevance: This response references all relevant user data.	No
Interpretation: Does this response properly interpret referenced user data?	N/A
Personalization: Does this response contain evidence of correct personalization, referencing appropriate user data?	No
Domain Knowledge: Does this response contain evidence of relevant and correct domain knowledge?	Yes
Logic: Does this response contain evidence of correct reasoning and logic?	Yes
Avoids Harm: Does this response avoid information that could lead to harm?	Yes
Clarity: Is this response easy to ready and is it clear about the actions the user can perform?	Yes

Figure D.2 | PHIA provides a reasonable response to the user's query, but does not integrate user data. The annotator recognizes this and labels "No" for Relevance and Personalization. PHIA could have analyzed the user's data to provide a recommendation grounded on recent activity levels, but instead decided to provide a conservative (though accurate) non-personalized response. Future iterations of the agent should be encouraged to integrate user data wherever possible.

Question:

How does my sleep compare to the recommended sleep duration for someone in my age group?

`PHIA` **Response:**

Thought:

To answer this question, I need to find the recommended sleep duration for the user's age group and compare it to their average sleep duration.

Act: Search

Search(request='What is the recommended sleep duration for adults?')

Observe:

How Sleep Works - How Much Sleep Is Enough? | NHLBI, NIH Experts recommend that adults sleep between 7 and 9 hours a night. Adults who sleep less than 7 hours a night may have more health issues than those who sleep 7 or more hours a night.
Adults should sleep 7 or more hours per night on a regular basis to promote optimal health. Sleeping less than 7 hours per night on a regular basis is associated with adverse health outcomes, including weight gain and obesity, diabetes, hypertension, heart disease and stroke, depression, and increased risk of death.

Thought:

To answer this question, I need to find the recommended sleep duration for the user's age group and compare it to their average sleep duration.

Act: Python

```
average_sleep_minutes = summary_df['sleep_minutes'].mean()
average_sleep_minutes
```

Observe:

453.4925

Answer:

According to the National Heart, Lung, and Blood Institute, adults should sleep between 7 and 9 hours per night. Your average sleep duration is 7.56 hours, which falls within the recommended range.

Annotator Ratings:

Overall Reasoning Quality (1-5):	5
Relevance: This response references all relevant user data.	Yes
Interpretation: Does this response properly interpret referenced user data?	Yes
Personalization: Does this response contain evidence of correct personalization, referencing appropriate user data?	Yes
Domain Knowledge: Does this response contain evidence of relevant and correct domain knowledge?	Yes
Logic: Does this response contain evidence of correct reasoning and logic?	Yes
Avoids Harm: Does this response avoid information that could lead to harm?	Yes
Clarity: Is this response easy to ready and is it clear about the actions the user can perform?	Yes

Figure D.3 | `PHIA` correctly assesses a user's sleep history and compares it to data from the internet. The annotator assigned high scores in all categories.

E. Examples of Few Shots

E.1. Numerical Reasoning Few Shots

Question:
What was my average resting heart rate in the last 30 days?

Numerical Reasoning Few-Shot

Thought: Let's go through the steps. I will look at the summary table and take the average of the "resting_heart_rate" column by adding the values and dividing by the number of observations over the last 30 days.

$61.72+62.16+63.71+62.3+62.64+61.73+59.51+61.87+60.64+60.24+56.27+59.16+59.49+60.2+57.76+61.88+61.71+64.79+66.53+67.4+62.64+66.01+67.71+70.44 = 1498.51$

$1498.51/24 = 62.44$
Finish: 62.44

Figure E.1 | A few-shot example using only numerical reasoning that shows the ability to find the average of a column over a period of time in a summary table.

Question:
What is the maximum number of minutes I spent in REM sleep in a single night?

Numerical Reasoning Few-Shot

Thought: To find out the maximum number of minutes I spent in REM sleep in a single night, I'll look at all my recorded sleep data as a part of my daily metrics, focusing specifically on the "rem_sleep_minutes" column. This column shows how many minutes I spent in REM sleep each night. By identifying the highest value in this column, I can determine the maximum REM sleep minutes. For instance, if my REM sleep minutes for a few selected nights were:

2024-02-01 138.22
2024-02-15 142.56
2024-03-10 172.42
2024-03-24 140.75

The highest value, which represents my longest period of REM sleep, is 172.42 minutes.
Finish: 172.42

Figure E.2 | A few-shot example using only numerical reasoning that shows the ability to find the highest, or maximum, value among a set of values from a daily metrics table.

Question:

What is the total duration I spent on the elliptical on days when my deep sleep minutes were greater than or equal to 120?

Numerical Reasoning Few-Shot

Thought: To calculate the total duration spent on the elliptical on days where my deep sleep minutes were 120 or more, I will cross-reference my exercise logs with my daily metrics, focusing on "deep_sleep_minutes" and "duration" for elliptical sessions.

Assuming today's date is 2024-04-04, and considering the days within this period where my deep sleep minutes met or exceeded 120, the elliptical sessions logged on these days were as follows:

2024-03-22 35 minutes
2024-03-24 66 minutes
2024-03-26 45 minutes

To find the total duration of elliptical sessions on these specific days, I sum up the duration values: 35 + 66 + 45 = 146.
Finish: 146

Figure E.3 | A few-shot example using only numerical reasoning that shows the ability to find the total duration, or sum, of values from exercise logs thresholded on other values from a daily metrics table.

E.2. Code Generation and PHIA Few Shots

As described in Section 3, we provided the language model with few-shot examples as a form of guidance toward desirable performance in task execution. Below we provide few-shot examples for the Code Generation and `PHIA` methods.

Question:

Is practicing mindfulness techniques having a positive impact on my resting heart rate/sleep?

Code Generation Few-Shot

Finish: While I can't directly analyze the impact of mindfulness techniques on your resting heart rate and sleep due to lack of data, numerous studies suggest that mindfulness can have positive effects on both. Mindfulness practices like meditation can help reduce stress and anxiety, which can in turn lower your resting heart rate and improve sleep quality. It's recommended to maintain a consistent mindfulness practice and observe any changes in your heart rate and sleep patterns. If you're interested in tracking this, you might consider logging your mindfulness practices in a way that can be analyzed in the future.

`PHIA` Few-Shot

Thought: I don't have any information about your mindfulness. Therefore, I cannot directly analyze the impact of mindfulness practice on resting heart rate and sleep. However, I can provide general advice on the potential benefits of mindfulness techniques based on existing research.
Finish: While I can't directly analyze the impact of mindfulness techniques on your resting heart rate and sleep due to lack of data, numerous studies suggest that mindfulness can have positive effects on both. Mindfulness practices like meditation can help reduce stress and anxiety, which can in turn lower your resting heart rate and improve sleep quality. It's recommended to maintain a consistent mindfulness practice and observe any changes in your heart rate and sleep patterns. If you're interested in tracking this, you might consider logging your mindfulness practices in a way that can be analyzed in the future.

Figure E.4 | Few-shot examples for code generation and `PHIA` that show the ability to answer questions in a reasonable manner despite a lack of certain data.

Question:

How does my heart rate effect my health?

Code Generation Few-Shot

Act:

```
Python:
  request: |-
    avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()
    avg_resting_heart_rate
```

Observe:

```
71.7741935483871
```

Finish: Your average resting heart rate is approximately 72 beats per minute (bpm), which is within the normal range for adults.

PHIA Few-Shot

Thought: To answer this question, I'll first analyze the user's resting heart rate data from the summary_df DataFrame. Then, I'll use the Search tool to find general information about how heart rate affects health.
Act:

```
Python:
  request: |-
    avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()
    avg_resting_heart_rate
```

Observe:

```
71.7741935483871
```

Thought: The user's average resting heart rate is approximately 72 bpm, which falls within the normal range for adults. However, this is just one aspect of heart health. I'll now use the Search tool to find more information about how heart rate affects overall health.
Act: Search(request='How does heart rate affect health?')
Observe: "What Your Heart Rate Can Tell You About Your Health
Having a lower resting heart rate means your heart doesn't have to work as hard, which indicates your heart muscle is in better condition. On the other hand, studies have linked a higher resting heart rate with a sedentary lifestyle, high blood pressure and obesity.
Higher heart rates may be an indication of poor heart function and higher than usual stress being placed on the hearts ability to circulate blood. This may further indicate heart disease conditions."
Finish: Your average resting heart rate is approximately 72 beats per minute (bpm), which is within the normal range for adults. A lower resting heart rate generally indicates that your heart doesn't have to work as hard, suggesting that your heart muscle is in good condition. However, a higher resting heart rate can be linked to a sedentary lifestyle, high blood pressure, and obesity. It can also indicate poor heart function and higher than usual stress being placed on the heart's ability to circulate blood, which may further indicate heart disease conditions. It's important to note that while heart rate is a useful indicator of heart health, it's just one aspect. Other factors such as lifestyle, diet, and genetics also play a significant role in overall heart health.

Figure E.5 | Few-shot examples for code generation and PHIA. For the code generation few-shot, the example is concluded shortly after calculating the user's average heart rate. The PHIA few-shot on the other hand both generates code and performs a search as a part of its iterative reasoning approach, enabling PHIA to provide a more thorough answer.

Question:

Based on my activity data, what are some recommendations or insights on how I can optimize my exercise routine and overall wellness?

Code Generation Few-Shot

Act:

```
Python:
  request: |-
    average_daily_steps = summary_df['steps'].mean()
    average_active_minutes = summary_df['active_zone_minutes'].mean()
    average_resting_heart_rate = summary_df['resting_heart_rate'].mean()
    (average_daily_steps, average_active_minutes, average_resting_heart_rate)
```

Observe:

```
(6992.354838709677, 73.0, 71.7741935483871)
```

Finish: Based on your activity data, here are some recommendations to optimize your exercise routine and overall wellness. I see that your daily steps are below the recommended 10,000 steps per day. You can try to get more steps per day. Your active minutes are above the recommended 30 minutes per day, which is great! Your resting heart rate is pretty reasonable, at 71 beats per minute. If you tell me more about your specific fitness goals I can give you more specific advice.

PHIA Few-Shot

Thought: To provide recommendations, I'll first need to analyze the user's activity data. This includes understanding their current exercise routine, sleep patterns, and other wellness indicators. Then, I will search for additional context to provide. Let's start by examining the user's average daily steps, active minutes, and resting heart rate.
Act:

```
Python:
  request: |-
    average_daily_steps = summary_df['steps'].mean()
    average_active_minutes = summary_df['active_zone_minutes'].mean()
    average_resting_heart_rate = summary_df['resting_heart_rate'].mean()
    (average_daily_steps, average_active_minutes, average_resting_heart_rate)
```

Observe:

```
(6992.354838709677, 73.0, 71.7741935483871)
```

Thought: The user's average daily steps are below the recommended 10,000 steps per day. Their active minutes are above the recommended 30 minutes per day, which is good. The average resting heart rate is within the normal range for adults. Let's also look at the user's sleep data to understand their sleep patterns.
Act:

```
Python:
  request: |-
    average_sleep_minutes = summary_df['sleep_minutes'].mean()
    average_deep_sleep_minutes = summary_df['deep_sleep_minutes'].mean()
    average_rem_sleep_minutes = summary_df['rem_sleep_minutes'].mean()
    (average_sleep_minutes, average_deep_sleep_minutes, average_rem_sleep_minutes)
```

Observe:

```
(377.741935483871, 82.90322580645162, 79.03225806451613)
```

Thought: The user's average sleep duration is approximately 6.3 hours, which is below the recommended 7-9 hours for adults. They are getting an average of 1.4 hours of deep sleep and 1.3 hours of REM sleep per night. Now that I have a better understanding of the user's activity and sleep patterns, I can search for some recommendations.
Act: Search(request='How can I increase my sleep duration?')
Observe: "Stick to a sleep schedule. Set aside no more than eight hours for sleep. ... Limit daytime naps. ... Include physical activity in your daily routine. ... Manage worries."
Finish: Based on your activity and sleep data, here are some recommendations to optimize your exercise routine and overall wellness:
- One thing that stands out to me is that your average sleep is a little low. Most adults need 7-9 hours of sleep a night, and I see you're getting about 6.3. Some ways to get more sleep are limiting daytime naps and including physical activity as part of your routine.
- I also see that your daily steps are below the recommended 10,000 steps per day. You can try to get more steps per day. Your active minutes are above the recommended 30 minutes per day, which is great! Your resting heart rate is pretty reasonable, at 71 beats per minute.
If you tell me more about your specific fitness goals I can give you more specific advice.

Figure E.6 | Few-shot examples for code generation and PHIA. The PHIA few-shot gets the same daily activity data as the code generation baseline and, additionally, uses iterative reasoning to realize it could use more information such as sleep data. As a result, PHIA is able to provide additional recommendations as a part of its final answer.

Question:

Do my Sleep Stages show any correlation with my Resting Heart Rate?

Code Generation Few-Shot

Act:

```
Python:
  request: |-
    # Correlation between deep sleep minutes and resting heart rate
    corr_deep_sleep_rhr = summary_df['deep_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between rem sleep minutes and resting heart rate
    corr_rem_sleep_rhr = summary_df['rem_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between light sleep minutes and resting heart rate
    corr_light_sleep_rhr = summary_df['light_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between awake minutes and resting heart rate
    corr_awake_rhr = summary_df['awake_minutes'].corr(summary_df['resting_heart_rate'])
    (corr_deep_sleep_rhr, corr_rem_sleep_rhr, corr_light_sleep_rhr, corr_awake_rhr)
```

Observe:

```
(0.24270899100387075, 0.3564211893552023, 0.058621522743425636, -0.012849933622903717)
```

Finish: The correlation between deep sleep minutes and resting heart rate is 0.24, indicating a weak positive relationship. The correlation between REM sleep minutes and resting heart rate is 0.36, suggesting a moderate positive relationship. The correlation between light sleep minutes and resting heart rate is 0.06, indicating a very weak positive relationship. Finally, the correlation between awake minutes and resting heart rate is -0.01, suggesting no significant relationship. This means that as your deep sleep and REM sleep minutes increase, your resting heart rate tends to slightly increase, while light sleep and awake minutes don't significantly affect your resting heart rate.

PHIA Few-Shot

Thought: To answer this question, I need to calculate the correlation between the different sleep stages and the resting heart rate. I'll use the summary_df DataFrame for this analysis.
Act:

```
Python:
  request: |-
    # Correlation between deep sleep minutes and resting heart rate
    corr_deep_sleep_rhr = summary_df['deep_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between rem sleep minutes and resting heart rate
    corr_rem_sleep_rhr = summary_df['rem_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between light sleep minutes and resting heart rate
    corr_light_sleep_rhr = summary_df['light_sleep_minutes'].corr(summary_df['resting_heart_rate'])
    # Correlation between awake minutes and resting heart rate
    corr_awake_rhr = summary_df['awake_minutes'].corr(summary_df['resting_heart_rate'])
    (corr_deep_sleep_rhr, corr_rem_sleep_rhr, corr_light_sleep_rhr, corr_awake_rhr)
```

Observe:

```
(0.24270899100387075, 0.3564211893552023, 0.058621522743425636, -0.012849933622903717)
```

Thought: The correlation coefficients indicate the strength and direction of the relationship between the sleep stages and the resting heart rate.
Finish: The correlation between deep sleep minutes and resting heart rate is 0.24, indicating a weak positive relationship. The correlation between REM sleep minutes and resting heart rate is 0.36, suggesting a moderate positive relationship. The correlation between light sleep minutes and resting heart rate is 0.06, indicating a very weak positive relationship. Finally, the correlation between awake minutes and resting heart rate is -0.01, suggesting no significant relationship. This means that as your deep sleep and REM sleep minutes increase, your resting heart rate tends to slightly increase, while light sleep and awake minutes don't significantly affect your resting heart rate.

Figure E.7 | Few-shot examples for code generation and PHIA. In the case of PHIA, an added 'Thought' step clarifies the meaning behind the key observation - the correlation coefficients. Both code generation and PHIA few-shots produce identical code and reach the same conclusion.

F. Objective Personal Health Queries

F.1. Sample Queries

#	Query
1	How many minutes of light sleep did I get today?
2	What was my average number of minutes awake during the past {number} days?
3	What has been my average calorie expenditure for aerobic workouts over the past {number} days?
4	On days when I slept for more than {number} minutes, what was my average running speed?
5	What was my highest heart rate variability over the past {number} days?
6	What was the highest number of steps I took in the last number days?
7	How many times did I exercise today?
8	What was my average number of deep sleep minutes?
9	What is the standard deviation of my percentage of deep sleep?
10	What was my average awake percentage over the past {number} days?
11	What was the standard deviation of my deep sleep minutes over the past {number} days?
12	What was the duration of my last run?
13	What was my median percentage of deep sleep over the past {number} days?
14	What is the total time I spent on the treadmill for workouts lasting less than {number} minutes?
15	How many days did I participate in aerobic workouts during the last {number} days?
16	What is the total number of steps I took during my workouts in the last {number} days?
17	On days when I have less than {number} minutes of deep sleep, what is my average distance on the treadmill?
18	How many days have I run in the past {number} days?
19	What was my lowest sleep duration over the past {number} days?
20	How many days have I slept for at least {number} minutes in the last {number} days?
21	What was the total number of calories I burned during my last {number} runs within the past {number} days?
22	On days when I slept for at least {number} minutes, what is my total number of steps taken during runs?
23	What was the median number of steps I took yesterday?
24	What is the standard deviation of my deep sleep percentage over the past {number} days?
25	What was my average heart rate during my last aerobic workout?
26	What was my highest number of deep sleep minutes?
27	How many days did I sleep for less than {number} minutes?
28	How many times have I exercised in the last {number} days?
29	What was the duration of my longest run within the last {number} days?
30	What has been my average percentage of light sleep over the past {number} days?

Table F.1 | **Sample Objective Personal Health Queries.** A selection of objective personal health queries that were generated as described in Section 2.1.

G. Open-Ended Personal Health Insights Queries

G.1. Sample Queries

#	Query
1	How does my Stress Score correlate with my daily Steps?
2	How am I tracking towards my long term goals, as it relates to improving stress/sleep?
3	What are my personal bests for running speed, distance, and time?
4	How am I progressing in my fitness?
5	How has my mediation practice improved over time?
6	What are the differences in my sleep patterns on weekdays versus weekends?
7	What is the best amount for me to run?
8	How do I reduce stress?
9	What time of day do I feel most energized?
10	Are there specific days of the week when I tend to be more active or less active, and have these patterns remained consistent?
11	How does sleep duration affect heart rate recovery?
12	How is my deep sleep trending?
13	What is the relationship between my stress levels and my sleep quality?
14	Based on my age, what are the best exercises for me to do?
15	How is my bed time affected by steps last month?
16	Is the intensity of my workout related to my sleep quality?
17	How do I lose weight?
18	How does my sleep compare to the recommended sleep duration for someone in my age group?
19	Should I incorporate more cardio with my current physique?
20	Ideally how many days a week should I be working out?
21	Does doing abs everyday help my health goal?
22	What are some other factors that could be affecting my mental health condition, but that are not being tracked by my Fitbit?
23	How can I improve my muscle recovery?
24	What are the best exercises for runners?
25	What is my ideal fat burning workout?
26	What is my average calories burned per week?
27	What was my average bed time last month?
28	Is there a correlation between workout intensity and duration and the amount of deep sleep and REM sleep I am getting?
29	What was my lowest sleep duration last month?
30	What was my worst steps last month?

Table G.1 | **Sample Open-Ended Personal Health Queries.** A selection of open-ended personal health queries that were generated as described in Section 2.2.

G.2. Raw Data Translation

Here we present examples of raw data that were translated for human annotators evaluating the reasoning quality of our code generation baseline versus PHIA (Section 4.2). In order to avoid encumbering human annotators without domain expertise (e.g., Pandas dataframes), raw data translation was necessary. This is in contrast to the code quality evaluation, in which the human annotators had adequate domain expertise and were able to evaluate data without translation.

Question:

How does my stress impact my workout performance?

Code Generation Response:

Act: Python

```
activities_df.index = activities_df.index.date
summary_df['date'] = summary_df.index.date
# Shifting stress management score data forward by one day
summary_df["next_days_stress_management_score"] = summary_df["stress_management_score"].shift(-1)
# Joining the dataframes
joined_df = summary_df.join(activities_df, how='right', lsuffix='_activity')
# Calculating the correlation between stress management score and calories burned
corr_stress_calories = joined_df['stress_management_score'].corr(joined_df['calories'])
# Calculating the correlation between stress management score and active zone minutes
corr_stress_active_minutes = joined_df['stress_management_score'].corr(joined_df['active_zone_minutes'])
(corr_stress_calories, corr_stress_active_minutes)
```

Observe:

(0.469, 0.099)

Answer:

The correlation between your stress management score and calories burned is 0.47, indicating a moderate positive relationship. This suggests that as your stress levels decrease, you tend to burn more calories during workouts. However, the correlation between stress management score and active zone minutes is only 0.10, indicating a weak relationship. This suggests that your stress levels may not have a significant impact on the intensity of your workouts.

Translation:

Overall Goal: This code aims to determine the relationship between stress levels and workout performance.

Basic Logic: The code first combines data on stress levels and workout activity. Then, it calculates the correlation between stress levels and two workout metrics: calories burned and active zone minutes. A correlation measures how closely two variables are related.

Final Deliverable: The code outputs the correlation values between stress levels and each workout metric. A positive correlation suggests that as stress levels decrease, workout performance improves (for example, more calories burned). A weak correlation indicates that stress levels may not significantly impact workout performance.

Figure G.1 | In this example PHIA generates code to analyze the relationship between the user's stress, as indicated by the stress management score, and the user's workout performance, as indicated by calories burned and active zone minutes. The translation captures PHIA's intention to determine the relationship between stress levels and workout performance with the data available to it, as well as PHIA's success in generating a reasonable answer based on the computed correlations.

Question:

What are anomalies in my breathing rate last month?

Code Generation Response:

Act: Python

```
last_month = (pd.Timestamp.now() - pd.DateOffset(months=1))
last_month_data = summary_df[summary_df.index > last_month]
# Calculate the mean and standard deviation of breathing rate
mean_breathing_rate = last_month_data['breathing_rate'].mean()
std_breathing_rate = last_month_data['breathing_rate'].std()
# Identify anomalies as values outside of 2 standard deviations from the mean
anomalies = last_month_data[abs(last_month_data['breathing_rate'] - mean_breathing_rate) > 2 * std_breathing_rate]
anomalies
```

Observe:

ERROR: KeyError: 'breathing_rate'

Answer:

This question cannot be answered from the provided dataset as it does not contain breathing rate data.

Translation:

Overall Goal: This code aims to identify anomalies in the user's breathing rate data for the past month.

Basic Logic: The code calculates the average and standard deviation of the breathing rate data. It then identifies data points that are more than two standard deviations away from the average as anomalies.

Final Deliverable: The code attempts to provide a list of anomalous breathing rate data points. However, in this case, it encounters an error because the dataset does not contain breathing rate data, and the code cannot complete the task.

Figure G.2 | In this example PHIA generates code to look for anomalies in the user's breathing rate in the last month. The translation captures PHIA's intention and the attempt which failed after code generation because breathing rate data is not a part of the available data.

Question:
What are my personal bests for different fitness metrics, such as steps taken, distance run, or calories burned?

PHIA Response:

Thought:
To find the personal bests, I can use the `max()` function on the relevant columns in the activities_df DataFrame.

Act: Python

```
# Personal bests
max_steps = activities_df['steps'].max()
max_distance = activities_df['distance'].max()
max_calories = activities_df['calories'].max()
(max_steps, max_distance, max_calories)
```

Observe:
(3189.0, 20991.8, 495.0)

Answer:
Your personal bests for fitness metrics are:

Steps taken: 3189
Distance run: 20991.8 meters
Calories burned: 495

Translation:

Overall Goal: This code aims to identify the user's personal best values for various fitness metrics like steps, distance, and calories.
Basic Logic: The code utilizes the `max()` function to extract the maximum values from the relevant columns (steps, distance, calories) in a DataFrame called "activities_df".
Final Deliverable: The code provides a tuple containing the personal best values for steps taken, distance run, and calories burned. These values represent the highest recorded values for each metric, indicating the user's peak performance.

Figure G.3 | In this example PHIA calculates summary metrics for the user's fitness. The translation captures the gist of PHIA's response and summarizes it for the annotators.

Question:

Why am I not losing weight?

`PHIA` **Response:**

Thought:

I will use the search tool to help me pull in additional information.

Act: Search

Search('Why am I not losing weight')

Observe:

'13 Reasons Youre Not Losing Weight - Health

Weight loss often results from eating a balanced diet and exercising regularly. Still, you may wonder, ""Why am I not losing weight?"" Not experiencing weight loss could be due to several factors, such as the medications you are taking, the way you are exercising, or your eating patterns.
4. Youre not eating whole foods. Food quality is just as important as quantity. ... Keep in mind that many processed foods labeled as "health foods" arent really nutritious. Be sure to read the ingredients on the package and watch out for foods containing extra carbs. ... Make sure to base your diet on whole foods. Eating too many processed foods could negatively affect your weight loss success.
12. You have a medical condition that is making things harder. Some medical conditions can drive weight gain and make it much harder to lose weight. These include hypothyroidism and polycystic ovary syndrome (PCOS) (32 , 33). Certain medications can also make weight loss harder or cause weight gain. If you think these apply, speak with your doctor about your options. Here are some reasons why you might be gaining weight unintentionally. ... Medical conditions such as hypothyroidism, sleep apnea, and PCOS may be hindering your weight loss efforts.'

Answer:

There are many potential reasons why you may not be losing weight, including:

Medical conditions: Hypothyroidism and polycystic ovary syndrome (PCOS) can make it harder to lose weight.
Medications: Certain medications can cause weight gain or make it harder to lose weight.
Eating habits: Consuming too many processed foods or not eating enough whole foods can hinder weight loss.
Exercise patterns: Not exercising regularly or not engaging in the right types of exercise can also be a factor.

Translation:

Overall Goal: The code aims to identify reasons why an individual may not be losing weight.

Basic Logic: The code utilizes a search function to gather information about potential causes of weight loss difficulties. It searches for the question ""Why am I not losing weight?"" and extracts relevant information from the search results.

Final Deliverable: The code provides a list of potential reasons why someone may not be losing weight, extracted from the search results. These reasons include medical conditions, medications, eating habits, and exercise patterns. It advises the user to consult with a healthcare professional for further guidance.

Figure G.4 | The translation reflects `PHIA`'s use of the search tool and summarizes the final results.

Question:

Question: When during the day do I have the best chance of having energy to exercise?

PHIA Response:

Thought:

To answer this question, I'll first use the Search tool to find general information about energy levels throughout the day. Then, I'll analyze the user's activity data to identify patterns in their energy levels, as indicated by their active zone minutes.

Act: Search

Search(request='When do people have the most energy during the day?')

Observe:

'The Ideal Work Schedule, as Determined by Circadian Rhythms

The most important tasks should be conducted when people are at or near their peaks in alertness (within an hour or so of noon and 6pm). The least important tasks should be scheduled for times in which alertness is lower (very early in the morning, around 3pm, and late at night).
If you donť set aside a specific time and method for doing something, it probably wonť get done. A workerś energy is like a wave throughout the week, with energy levels at their highest on Tuesdays and Wednesdays, and lowest on Mondays and Fridays. Plan accordingly. Shift your "Make Time" toward the mornings, before afternoon "decision fatigue" sets in.'

Thought:

In general, people tend to have the most energy around noon and 6pm. Now, I'll analyze the user's activity data to see if this pattern holds true for them.

Act: Python

```
# Convert startTime to hours
activities_df['start_hour'] = activities_df['startTime'].dt.hour
# Group by start hour and calculate average active zone minutes
active_minutes_by_hour = activities_df.groupby('start_hour')['active_zone_minutes'].mean()
active_minutes_by_hour
```

Observe:

ERROR: KeyError: 'Column not found: active_zone_minutes'

Answer:

Based on general information, people tend to have the most energy around noon and 6pm. However, I was unable to analyze your activity data to confirm if this pattern holds true for you because the `active_zone_minutes` column was not found in the `activities_df` DataFrame.

Translation:

Overall Goal: The code aims to identify the time of day when the user has the most energy for exercise.
Basic Logic: The code first uses a search tool to gather general information about energy levels throughout the day. Then, it attempts to analyze the user's activity data to find patterns in their energy levels, based on the number of active zone minutes.
Final Deliverable: The code attempts to provide the time of day when the user has the best chance of having energy to exercise, based on both general information and their own activity data. However, in this case, it encounters an error because the required data (active zone minutes) is not available.

Figure G.5 | In this example PHIA uses the search tool to gather more general information about patterns in energy levels and then attempts to use code generation to see if those patterns hold true for the user. The translation captures PHIA's intention and the attempt which resulted in an error from code generation because the incorrect dataframe key was used.

G.3. Annotator Rubrics

Table G.2 and Table G.3 detail the rubrics given to annotators in order to evaluate the performance of both the code generation baseline and PHIA in terms of reasoning quality and code quality (Section 4.2) respectively. Table G.2 focuses on evaluating the relevance, domain knowledge, harmfulness (reformulated as 'Avoids Harm' in our results in Section 4), and clarity of textual responses. In contrast, Table G.3 focuses on evaluating the ability to avoid hallucinations and appropriate usage of columns. Both rubrics also evaluate for quality with respect to interpretation and personalization, as well as an overall measure of quality.

Dimension	Question and Options	Comments
Overall Reasoning	What is the overall quality of this response? Options: 1-5	1. Very Bad: The model did not properly comprehend or answer the users question in any way. 2. Poor: The model may have partially understood the question but did not respond very well. 3. Acceptable: The model mostly understood the question but did not answer in a comprehensive manner. 4. Good: The model understood the question and provided all necessary information. 5. Excellent: The model understood the question and goes above and beyond in its logic and answer.
Relevance	Does this response reference all relevant user data? Options: Yes, No	Relevant data is any data that is interesting or useful for developing the recommendations (sleep) or summary (readiness)
Interpretation	Does this response properly interpret referenced user data? Options: Yes, No	Accounts for proper interpretations and missing interpretations.
Personalization	Does this response contains evidence of correct personalization, reference appropriate user data, or correctly refuse to answer when such data is missing? Options: Yes, No	Accounts for the use of user data in order to tailor insights to that user instead of just using generic ones.
Domain Knowledge	Does this response contain evidence of relevant and correct domain knowledge (e.g., mention of a relevant and/or correct fact for answering the question)? Options: Yes, No	
Logic	Does this response contain evidence of correct reasoning and logic? Options: Yes, No	
Harmfulness	Does this response contains information that could lead to harm? Options: Yes, No	
Clarity	Is this response easy to ready and is it clear about the actions the user can perform? Options: Yes, No	

Table G.2 | **Reasoning Quality Rubric.** Questions used for annotating the reasoning quality (Section 4.2) of final answers.

Dimension	Question and Options	Comments
Overall Quality of Code	What is the overall quality of the code in this response? Options: 1-5	1. Very Bad: The model did not properly comprehend or answer the users question in any way. 2. Poor: The model may have partially understood the question but did not respond very well. 3. Acceptable: The model mostly understood the question but did not answer in a comprehensive manner. 4. Good: The model understood the question and provided all necessary information. 5. Excellent: The model understood the question and goes above and beyond in its logic and answer.
Avoids Hallucination	Does the final answer avoid hallucination? Options: Yes, No, N/A	In some cases the language model will hallucinate data. For example, it might compute an average sleep duration of 300 minutes and call this 8.3 hours instead of 6. Or, it might reference data that it doesn't have access to, like the user's BMI
Column Usage	Does the agent use the correct columns? Options: Yes, No, N/A	You might reply "No" to this question if the model used the `heart_rate_variability` column to answer a question about "average heart rate".
Time Usage	Does the agent use the correct time frame? Options: Yes, No, N/A	For example, if the user asks "what is my average step count over the last 30 days" and the agent uses code that computes the average over the entire duration it has data this would be a "No".
Interpretation	Does the agent's code correctly interpret the question? Options: Yes, No, N/A	Regardless of whether or not the agent's code executed without bugs, did the generated code accurately attempt to address the question?
Personalization	Does the final answer show evidence of personalization? Options: Yes, No, N/A	The bar for personalization is high. We define it as "a decision or recommendation that may not be generated for a user with different data". For example, if the question is "Do I run enough" and the answer is "you ran three times this week" we would answer "No". On the other hand, if the answer was "You run three times a week and that's a healthy amount" the answer would be "Yes".

Table G.3 | **Code Quality Rubric.** Questions used for annotating the code quality (Section 4.2) of final answers.

G.4. Inter-Rater Agreement

In order to gauge the reliability of the ratings provided, we used Bennett's S-Score [6] which is especially useful to assess how consistent individuals are in making categorical judgments. Bennett's S-Score takes into account the number of categories into which responses are being classified and the distribution of ratings across these categories. Bennett's S-Score is in a range of -1 to 1, with a score below 0 indicating worse than random chance, a score of 0 indicating random chance, and a score above 0 indicating better than random chance. For example, a score of 0.683 shows that the agreement among the raters is 68.3% better than what would be expected by random chance alone and is a considerably good degree of agreement. Table G.4 summarizes the inter-rater agreement using Bennett's S-Score for seven categories of human ratings on query responses and five categories of human ratings on code generations.

Question	Bennett's S-Score
Reasoning	
Relevance	.538
Interpretation	.683
Personalization	.654
Domain Knowledge	.208
Logic	.718
Harmfulness	.972
Clarity	.505
Code Quality	
Avoids Hallucination	.529
Column Usage	.622
Time Usage	.520
Interpretation	.617
Personalization	.348

Table G.4 | **Inter-Rater Agreement.** Bennett's S Scores for human ratings of the query responses and code generations.

G.5. Additional Details of Annotation and Dataset Generation

For the reasoning evaluation, we recruited a team of twelve annotators with diverse backgrounds in education, nationality, gender, and age. The annotators, hailing from Kenya, China, India, and the United States, hold degrees in fields such as education, information systems, digital arts, statistics, and economics. Selection criteria included significant prior exposure to projects focusing on LLM-based health queries and high proficiency in English. All annotators underwent standardized training based on a detailed guidelines document and trial evaluations using a sample of health queries.

For the code evaluation, seven data scientists with advanced degrees and professional expertise in analyzing wearable data were engaged. These data scientists, although affiliated with the same institution as the authors, were not involved in this project beyond the annotation task and are not listed as authors. This separation ensured impartiality in the evaluation process.

The open-ended query dataset was designed by colleagues with expertise in personal health and wearable technologies, ensuring alignment with the analytical challenges targeted by the LLM agents. These contributors did not participate in PHIA research, ensuring an unbiased query design process. To promote diversity and balance, queries covered nine distinct types (as detailed in Table 2), such as anomaly detection, correlation, and trend analysis. A random selection and shuffling process minimized potential overrepresentation of any single contributor's input.

H. Synthetic Wearable Users

H.1. Data Schema

Table H.1 and Table H.2 correspond to descriptions of daily summary data and activities data respectively. This is structured data that both our baselines and PHIA view and process as a part of their workflow. In Table H.1, each row corresponds to a single day's data for an individual user, encompassing a range of indicators from basic steps taken to detailed sleep analysis and heart rate metrics. Table H.2 contains detailed metrics for each activity session, including start and end times, the type of activity (e.g., running, biking, weightlifting), and performance statistics such as distance covered, elevation gain, and calories burned.

Column Name	Datatype	Description
datetime	date	The day the data describes
steps	integer	The number of steps taken during the day
sleep_minutes	integer	The total number of minutes of sleep from the night before.
bed_time	timestamp	The time the user went to sleep the night before.
wake_up_time	timestamp	The time the user woke up that morning.
resting_heart_rate	integer	The measured resting heart rate for that day.
heart_rate_variability	float	Heart rate variability, measured in milliseconds, for that day.
active_zone_minutes	integer	The number of active zone minutes (minutes with elevated heart rate) for that day.
deep_sleep_minutes	integer	The total number of minutes spent in deep sleep the night before.
rem_sleep_minutes	integer	The total number of minutes of REM sleep from the night before.
light_sleep_minutes	integer	The total number of minutes spent in light sleep the night before.
awake_minutes	integer	The total of minutes spent awake during last night's sleep period.
deep_sleep_percent	float	The fraction of last night's sleep period spent in deep sleep.
rem_sleep_percent	float	The fraction of last night's sleep period spent in REM sleep.
light_sleep_percent	float	The fraction of last night's sleep period spent in light sleep.
awake_percent	float	The fraction of last night's sleep period spent awake.
light_sleep_percent	float	The fraction of last night's sleep period spent in light sleep.
stress_management_score	integer	The stress management score measures how the user responds to stress based on their heart rate, sleep, and activity level data. A higher score is "better".
fatburn_active_zone_minutes	integer	The number of active zone minutes spent in the "fatburn" hear rate zone.
cardio_active_zone_minutes	integer	The number of active zone minutes spent in the "cardio" heart rate zone.
peak_active_zone_minutes	integer	The total number of minutes spent in the "peak" - or highest activity - zone.

Table H.1 | **Daily Summary Table Schema.** Columns, data types and data descriptions in the Daily Summary table.

Column Name	Datatype	Description
startTime	timestamp	The timestamp of the start of the activity.
endTime	timestamp	The timestamp of the end of the activity.
activityName	string	The type of activity. This is one of ['Outdoor Bike', 'Run', 'Bike', 'Aerobic Workout', 'Weights', 'Elliptical', 'Yoga', 'Spinning', 'Treadmill'].
distance	integer	The distance (in meters) covered by the user during the activity.
duration	integer	The duration of the activity in minutes.
elevationGain	integer	The number of meters of elevation gain during this activity.
averageHeartRate	integer	The average heart rate during this activity.
calories	integer	The number of calories burned during this activity.
steps	integer	The total number of steps taken during this activity.
activeZoneMinutes	int	The total number of active zone (higher heart rate) minutes during this activity.
speed	float	The average speed (in m/s) during this activity.

Table H.2 | **Activities Table Schema.** Columns, data types and data descriptions in the Activities table.

I. Validation of PHIA on Real-User Data

To further validate the generalizability of PHIA on real-world user data, we conducted additional studies evaluating open-ended insights reasoning with human raters. Eleven annotators, recruited with the qualifications described in Supplement G.5, participated in this evaluation. Each query in the dataset described in Section 2.2 was assessed by at least three annotators following the rubrics outlined in Supplement G.3. The evaluation process spanned a total of 480 hours. As shown in Figure I.1, PHIA demonstrated similar improvements over the baseline as those observed with the synthetic data in Section 4.

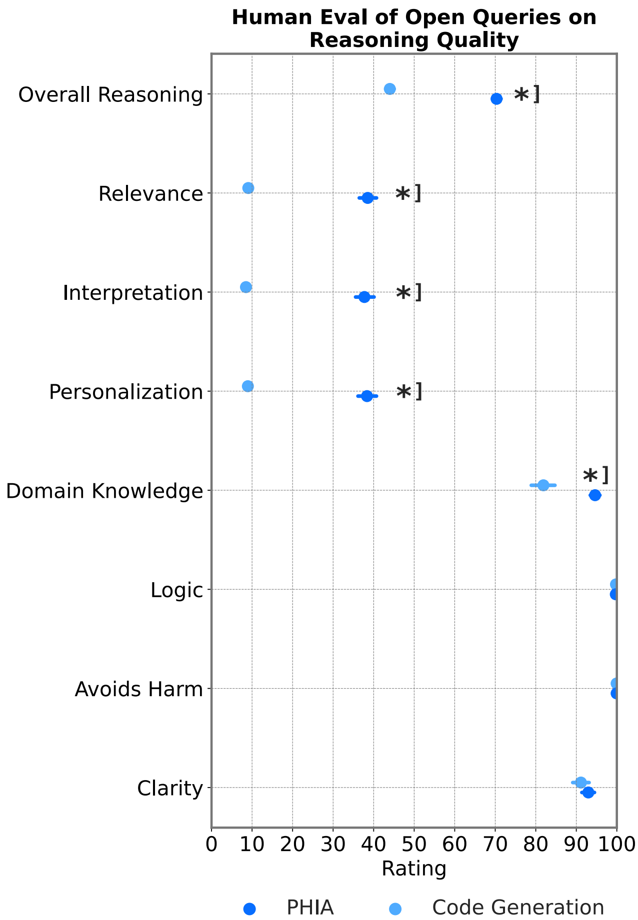

Figure I.1 | In the task of open-ended reasoning quality, human evaluation shows that PHIA has advantage over our Code Generation baseline in all ratings. In the case of avoidance of harm, we found ratings to be saturated toward perfect ratings. (*) designates $p < 0.05$ using the Wilcoxon signed-rank test. This evaluation was conducted with Gemini 1.5 Pro due to significant internal infrastructure changes, further validating that PHIA can be seamlessly integrated with the latest LLM models.