

---

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

---

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

## Abstract

Despite the proliferation of wearable health trackers and the importance of sleep and exercise to health, deriving actionable personalized insights from these data remains a challenge. The recent rise of large language models (LLMs) coupled with their integration into tools have unlocked capabilities in code generation and information retrieval. Yet, the application of these advanced models in analyzing wearable data is still largely untapped. In this paper, we introduce the Personal Health Insights Agent (PHIA), an agent system that leverages LLMs along with state-of-the-art code generation and information retrieval tools to analyze and interpret behavioral and health data from wearables. We demonstrate that PHIA can answer health insights questions across domain sub-populations and query types. We support this claim by curating two benchmark question-answering datasets of 4000+ health insights questions to quantitatively and qualitatively evaluate the capabilities of tabular and numerical reasoning for personal health. Our extensive qualitative and quantitative evaluation with expert human raters reveal that PHIA can accurately address over 80% of factual numerical questions and more than 83% of crowd-sourced open-ended questions, providing in-depth insights and interpretations. This work has implications for advancing behavioral health across the population, potentially enabling individuals to interpret their own wearable data, and paving the way for a new era of accessible, personalized wellness regimens that are informed by data-driven insights.

## 1 Introduction

The potential of large language models (LLMs) extends beyond the foundational capacity to generate language for complex tasks that require reasoning and decision-making [29]. In the domain of health, LLMs have the potential to increase access, efficiency and accuracy in tasks ranging from medical question-answering [21, 22, 27], medical education [7, 26], electronic health record analysis [9, 20, 30], interpretation of medical images and assessments [11, 27] to generating diagnoses [8, 14]. LLMs can be augmented with additional software tools extending their capabilities, examples of which include code generation [13] and factual information retrieval [12, 34]. Foundation models not only exhibit proficient natural language or image understanding and generation, but also “emergent” capabilities like reasoning, planning, and decision-making [29]. This confluence of advanced reasoning and generation capabilities has led to a rise in interest in LLM-based *agents* - an emerging class of artificial intelligence (AI) systems that use LLMs to interact with the world [25]. Their modular design, which encompasses information storage,

---

<sup>1</sup>\* Corresponding authors: mikeam@cs.washington.edu, xliucs@google.com. Mike A. Merrill is a PhD student at the University of Washington. Work done during an internship at Google.

action space and decision-making procedures as a conceptual framework, facilitates a more interpretable structure and a broader action space of planning with APIs and tools. In particular, the tool-use ability enables agents to dynamically interact with the environment and achieve versatile expertise [17, 19].

Most LLM-based health agents have been trained and evaluated on data collected from clinical sources [20]. However, the application of these agents to personal health and wellness data represents an unexplored area. These data are often derived from personal devices such as wearables, distinguished by its continuous and longitudinal measurement that captures high granularity observations of physiology and behavior in-situ rather than in a clinical setting. In practice, these datasets largely consist of multi-dimensional time-series data from body worn sensors. Research studies have highlighted the significant health impacts of physical activity and sleep patterns, emphasizing the potential for personalized health data [1, 2, 15, 23, 24]. 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 [24]. Those with frequent sleep disturbances were associated with an increase in risk of hypertension by 20% and cardiovascular diseases by 1.5 times [15].

Deriving intelligent responses and insights to personal health queries that involve analysis or interpretation of wearable data is non-trivial. This data is usually collected without expert supervision. As a result, people often do not have access to health professionals that could aid in data interpretation. For example, a common question from users of wearable devices may be “How can I get better sleep?”, but arriving at an ideal response would involve performing a series of complex, independent analytical steps across multiple irregularly sampled time series such as: calculating average sleep duration, identifying anomalies in the individual’s sleep pattern over a reasonable period of time for statistical precision (e.g., one month), contextualizing these findings within the broader spectrum of the individual’s health, integrating knowledge of population norms of sleep, and offering tailored sleep improvement recommendations. These steps not only involve numerical analysis but also the interpretation of what constitutes healthy sleep, taking into account the nuances of individual health profiles.

Until recently, it would have been optimistic to think that a machine learned model would be capable of all of these steps without substantial, custom engineering efforts. However, the advances in the capabilities of LLM agents enable this level of level of multi-step reasoning. As such, LLM agents represent a significant opportunity when it comes to deriving insight from personal health data. If an agent can autonomously decompose complex tasks, reason using both their internal knowledge base and external analytical tools, and generate actionable insights, then it could present real utility to individuals, and collectively population health. By incorporating time-series wearable data, they are uniquely positioned to transform complex numerical data into comprehensible health insights.

In this paper, we address the gap between personal health data analysis and agent-based systems by introducing the first LLM agent for personal health, called Personal Health Insights Agent (PHIA). PHIA incorporates advanced code generation, search engine integration, and the ReAct prompting technique [31] to facilitate multi-step reasoning and address thousands of real-world health queries. Our system is capable of navigating complex health-related questions, offering personalized health information and recommendations that were previously unattainable at such a scale and accuracy.

Specifically, our key contribution of this paper are:

- We introduce the first LLM-based health agent framework that utilizes code generation and web search tools for in-depth analysis of thousands of health insight questions.
- We conduct a 600-hour human evaluation of more than 6000 model responses with 19 raters and an automatic evaluation of 12000 model responses to demonstrate the superior capabilities of LLM agents in reasoning about time-series behavior health data as well as deep health insight interpretations, comparing against non-agent LLM-based code generation and text-only reasoning approaches.
- We will release the first set of high-fidelity synthetic wearable users, sampled from high-volume anonymized production data to ensure a rich representation of real-world variation.

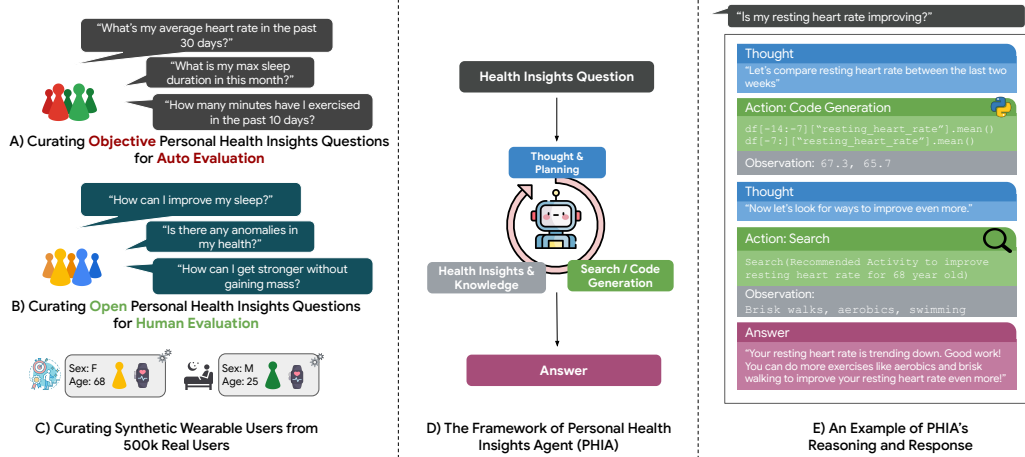


Figure 1: **Overview of Personal Health Insights Agent (PHIA)**. (A)-(C) illustrate the process of curating objective and open-ended health insight queries, along with the creation of synthetic wearable user data, which are utilized to evaluate PHIA's capabilities in reasoning and understanding health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

- We will release the first open-source, synthetic, personal health insights evaluation dataset, comprised of over 4000 closed and open-ended questions across multiple domains for both automated and human evaluation.

## 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 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 and understand personal health insights. The first dataset comprises solvable and objective health insights queries designed for automated evaluation. The second dataset consists of open-ended health insights queries intended for human evaluation. Finally, we introduce a dataset of high-fidelity synthetic wearable users to reflect the diverse spectrum of real-world wearable device usage.

### 2.1 Objective Personal Health Insights

**Definition.** Objective personal health queries are characterized by their 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 straightforwardly classified as correct or incorrect.

**Dataset Curation.** To generate objective personal health queries, we developed a framework aimed at the systematic creation and assessment of solvable personal health insights questions and their respective answers. This framework is based on manually crafted templates by a domain expert, providing a schema for the formulation of these queries. These templates are designed to incorporate a broad spectrum of variables, encompassing essential analytical functions, data metrics, and temporal definitions. This method enables the production of a varied collection of health insights queries, distinguished by their specificity and adaptability to the intricacies and demands of health data analytics.

Query Type	Count	Example
Correlation	40	How does my Stress Score correlate with my daily Steps?
Problematic	30	Does not eating make your stomach look better?
Query	19	What is the best amount for me to run?
General Knowledge	16	What's a good meal for breakfast, that will meet most of my nutritional needs for the day?
Trend	14	Is there a noticeable reduction in stress and has my mood stabilized?
Summary	11	How am I progressing in my fitness?
Time-Frame Comparison	9	What are my sleep patterns during different seasons?
Personal Best/Worst	8	What was my worst steps last month?
Compare vs Population	8	Is my resting heart rate of 52 healthy for my age?
Anomaly	7	If I noticed spikes in my data, what action should I take?
Average	5	What are my stats for the following...Average sustained heart rate while exercising? Resting heart rate? Top speed when running? Average/ top jump height?
Lowest/Highest	5	What is the max number of calories I burn?
Total	172	

Table 1: A summary of open-ended questions used in our human evaluation.

Consider the following scenario for demonstration: a template is established to calculate daily averages 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 is my average daily steps last week?" and the code `daily_metrics["steps"].during("last week").mean()` can be utilized to generate the corresponding response. It is worth noting that `during()` is a custom function to handle date interpretation in Python Pandas library. A total of 4000 health insights questions were generated using this approach. All of these questions were manually evaluated by a domain expert at the intersection of data science and health research to ensure their precision and comprehensibility.

## 2.2 Open-Ended Personal Health Insights

**Definition.** Open-ended 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 data analysis tools and understanding in wearable health data.

**Dataset Curation.** To curate a such complex dataset, a survey was conducted with a sample of the authors' colleagues, all of whom had relevant expertise in consumer health research and development, soliciting their potential inquiries for a hypothetical AI agent equipped with access to their personal wearable data. Participants were asked, "If you could pose questions to an AI agent that analyzes your smartwatch data, what would you inquire?" This survey generated approximately 3,000 health insights questions, which were subsequently manually categorized into one of twelve distinct query types (Table 1). For evaluation feasibility reasons, a smaller test dataset was created, comprising 200 randomly selected questions. From this subset, questions with high semantic similarities were excluded, resulting in a final tally of 172 distinct personal health queries. These were intentionally excluded from model development to avoid influencing the generation of few-shot examples.

## 2.3 Synthetic Wearable Users

**Definition.** To effectively evaluate both objective and open-ended personal health insights queries, high-fidelity wearable data is essential. To maintain the privacy of wearable device users, we developed the first synthetic data generator for wearable data. This generator is based on a large-scale anonymized dataset from 30000 real wearable users who agreed on contributing their datasets for research purposes. Each of the synthetic wearable users has two tables – one of daily activity statistics (e.g. sleep duration, bed time and total step count for each day) and another of describing discrete activity events (e.g. a 5km run on 2/4/24 at 1pm). The schema of these tables are available in Appendix H. Importantly, these data mirror the real data that are collected by consumer wearables like Fitbits.

**Dataset Curation.** To ensure representation of diverse user profiles, we aggregated data over a 31-day period, requiring a minimum of 10 days of data availability for inclusion. We utilized a Conditional Probabilistic Auto-Regressive (CPAR) neural network [16, 33], specifically designed to manage sequential multivariate and multi-sequence data, while integrating stable contextual attributes like age and gender. This approach distinguishes between unchanging context (i.e., non-time-varying 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.

## 3 The Personal Health Insights Agent (PHIA)

Language models are useful tools for autoregressive text generation but struggle to plan future actions, interpret outside information, or use tools on their own ([5, 28]). Accordingly, in order for an LM to perform complex time-series wearable data analysis to interpret personal health behavior, it must be embedded into a larger *agent framework* that interprets the LM's outputs and allows it to interact with the external world.

**Iterative & Interactive Reasoning.** PHIA is based on the widely recognized ReAct agent framework [32], where an "agent" refers to a system capable of performing actions autonomously and incorporating observations about these actions into decisions (Figure 1). 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 an action plan. This critical phase enables the model to assimilate all available information, thereby determining the most effective action for inching closer to its objective—formulating a response to the query at hand. Following this framework, in the *Act* stage, the language model implements its strategy by dispatching commands to one of its auxiliary tools. These tools, operating independently from the model, provide feedback to the agent's state by executing specific tasks. In PHIA, significant tools include a Python data analysis runtime for arithmetic operations and a web 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 scripts' results or relevant web pages sourced through the search API in this phase.

**Wearable Data Analysis with Code Generation.** During an *Act* stage, the agent engages with wearable tabular data through the creation of raw Python scripts executed within a customized sandbox runtime environment. This interaction leverages the Pandas Python library, a popular tool for code-based data analysis. The numerical results derived from code generation are factual, and precision in arithmetic calculations is reliably maintained. Moreover, this approach can help preserve the confidentiality of the user's 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 quality of reasoning process by integrating a web search based mechanism to retrieve the latest and relevant health information based on the query. This custom search capability systematically extracts and interprets content from leading search results. This approach presents dual benefits: it directly attributes

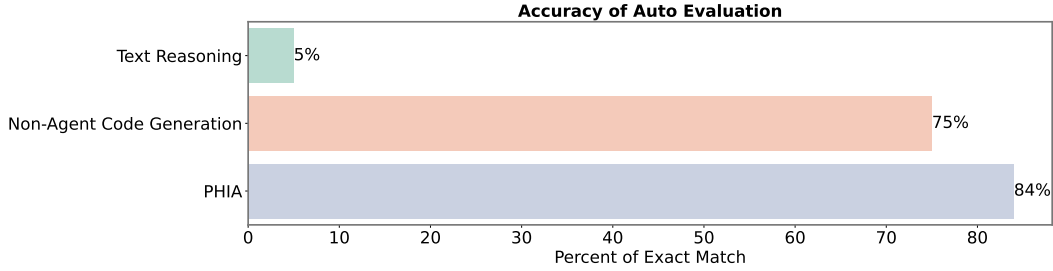


Figure 2: **Automatic Evaluation.** PHIA scores better than the Non-Agent code generation and standard LLM text-reasoning baselines on objective personal health insights queries presented in Section 2.1. Accuracy is based on an exact match to within two digits of precision.

information to its original authoritative 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. Such integration of external knowledge also helps diminish uncertainty in the model’s responses to health queries.

**Mastering Tool Use.** A popular technique for augmenting the performance of agents and language models is few-shot prompting [4]. This approach entails providing the model with a set of high-quality examples to guide it on the desired task without expensive and laborious fine-tuning. To create in-context examples that can guide model generations across a variety of topics, we initially calculate a sentence-T5 embedding [18], a form of vector representation capturing the essence of each query—for all entries in our validation dataset which includes more than 1000 handcrafted health insights queries. Subsequently, we apply K-means clustering, a method to group data points based on their similarity, on these embeddings, targeting 20 distinct clusters. We then select queries closest to the centroid of each cluster. For each chosen query, we carefully craft a ReAct trajectory that demonstrates how to produce a high quality response with iterative planning, code generation, and web search. Refer to Appendix ?? for more examples.

## 4 Experiments & Results

### 4.1 Baselines

To understand the necessity of tools (e.g., code generation, web search) and the agent framework, we construct two language model baselines to demonstrate PHIA’s performance.

**Text Reasoning.** Language models have demonstrated modest mathematical ability and so it may be the case that PHIA’s code interpreter is not necessary to answer personal health queries. In this methodology, user data is structured into a markdown table format and directly supplied to the language model, coupled with the corresponding query. Analogous to PHIA, we designed a set of few-shot examples to guide model to execute rudimentary operations such as calculating the mean of a data column.

**Non-Agent Code Generation.** Is it necessary to use an agent to deliver personal health insights? As a comparative benchmark, we introduce a non-agent code generation model which can only generate answers in a single inference step. In contrast to PHIA, this model lacks a reflective ‘Thought Step’ and is unable to strategize multiple steps ahead. This limitation hampers the model’s ability to perform comprehensive multi-step analysis as is feasible in PHIA. To enhance this baseline model, it is fortified with a unique set of few-shot examples that employ identical queries to those used in PHIA (Section 3), albeit with responses and code crafted by humans to mirror the restricted capabilities of the non-agent model.

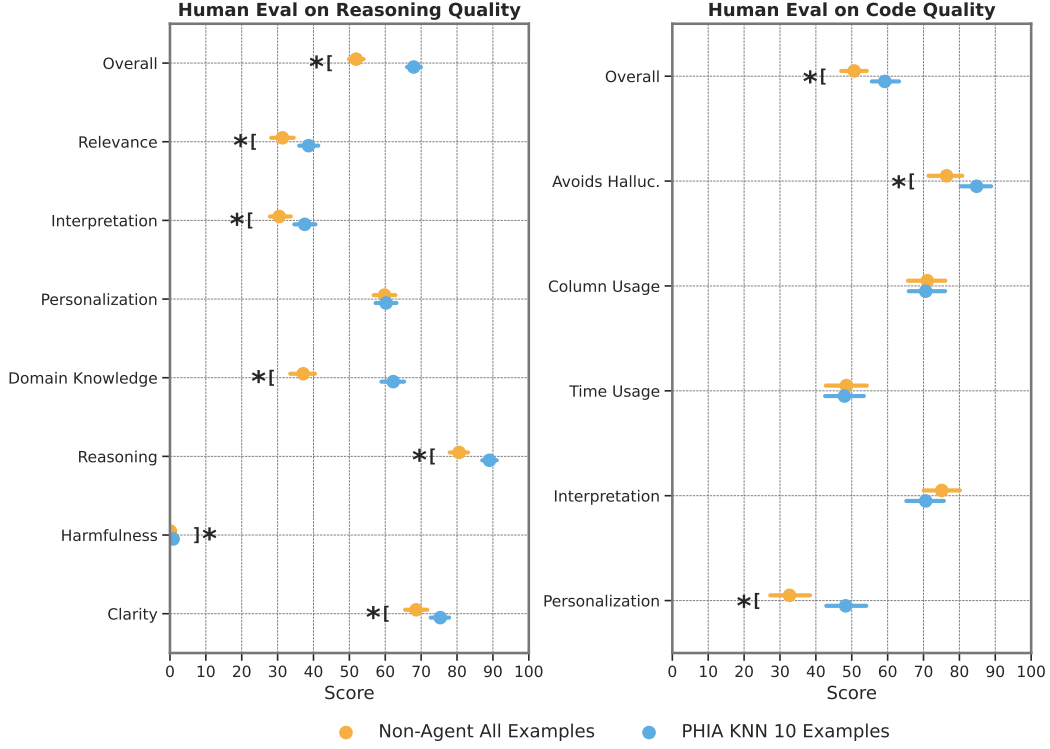


Figure 3: **Human Evaluation.** Our human evaluation on reasoning quality and code quality shows that PHIA has a significant advantage over the non-agent baseline in all ratings except personalization with respect to reasoning quality and column usage, time usage, and interpretation with respect to code quality. 'Non-Agent All Examples' refers to our non-agent baseline while 'PHIA KNN 10 Examples' refers to our agent baseline that has a cognitive phase ('Thought step'). (\*) designates  $p < 0.05$ .

## 4.2 Experiments

**Automatic Evaluation on Objective Queries.** Some personal health queries have tractable solutions that afford automatic evaluation as defined in Section 2.1. To study PHIA's performance on these questions, we evaluated PHIA and baselines on all 4000 objective insight questions in our objective personal health insights dataset. A question 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 wrong).

**Human Evaluation on Reasoning.** Numerous personal health questions require complex reasoning, which demands precise interpretation of queries to integrate user-specific data with relevant health guidelines and expert knowledge. To assess this **reasoning** capability, we recruited a team of twelve independent raters who had substantial familiarity with wearable data to evaluate the quality of reasoning of PHIA and our Non-Agent baseline in the open-ended personal health insights dataset defined in Section 2.2. Due to minimal experience with Python data analysis, the annotators were provided with model responses that had been reformulated by Gemini Ultra to replace Python code with explanatory text. Annotators were tasked with assessing whether each response demonstrated the following attributes: relevance of data utilized, accuracy in interpreting the question, personalization, incorporation of domain knowledge, correctness of the reasoning process, absence of harmful content, and clarity of communication. Additionally, they rated the overall quality of each response using a Likert scale ranging from one (lowest quality) to five (highest quality). All the evaluation questions were distributed in a

manner that ensured that each model response was rated by at least three unique annotators. Comprehensive rubrics and instructions for annotation can be found in Table 7. To standardize comparisons across different metrics, final scores were obtained from normalizing the original ratings on a scale of 1-5 into a range of 0-1 and then multiplying by 100. Subsequent categorical scores are the proportion of humans that responded 'Yes'. For example, an answer of 'Yes' for 'Domain Knowledge' would indicate that a human evaluator found the response to show an understanding of domain knowledge.

**Expert Evaluation on Code Quality.** To assess the quality of the code outputs of PHIA and our Non-Agent baseline, we recruited a team of seven experienced data scientists with extensive experience in analyzing data from wearable devices. Collectively, these annotators brought several decades of relevant experience to the task. We distributed the model responses from PHIA and the Non-Agent Baseline such that each sample was independently evaluated by three different annotators. Annotators were blinded to the experimental condition (i.e. which model produced the outputs). Unlike in the reasoning evaluation with raters who don't have experiences with Python data analysis, annotators were provided with the complete model response from each method, including generated python code, *Thought* steps, and any error messages. Annotators 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 Appendix F).

**Error Analysis.** While expert code annotators agree that PHIA's analysis is of overall higher quality than the Non-Agent baseline it is also possible to automatically detect and evaluate fatal runtime errors that cause the Python Data Analysis Runtime to crash. Examples include attempts to index columns that don't exist, or import libraries that are not accessible to the agent, or syntax mistakes. To test this, we calculate 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. Intuitively this number reflects how often a method fails to generate valid code.

### 4.3 How does PHIA Perform?

**PHIA correctly answers objective personal health queries.** PHIA outperforms the non-agent baseline by 13% across all personas (Figure 2). This shows that the additional complexity afforded by the agent framework helps the model perform better, even on simple tractable questions that require limited abstract reasoning. Notably, the *Text Reasoning* baseline scores below 7% accuracy on each persona, indicating that text-only reasoning is a poor method for answering personal health insights queries. We believe this is due to LLM's poor mathematical and tabular reasoning. Given this low performance we exclude this method from further evaluation.

**PHIA demonstrates superior reasoning on open-ended queries.** Overall, PHIA demonstrates a significant improvement over the Non-Agent baseline in all but one dimension (Figure 3). Most notably, overall quality was substantially higher for PHIA than the non-agent (3.82 versus 3.12 average Likert rating). Other significant improvements over baseline include the domain knowledge category (63 vs 38) and overall reasoning. Taken as a whole, this evaluation indicates that PHIA's agent-based method produces substantially higher quality reasoning than the non-agent baseline and is much more effective at addressing user-provided queries than its base language model alone.

**PHIA shows improved personal health data analysis abilities.** The results from expert evaluation indicate that PHIA improved over the Non-Agent code generation baseline in overall code quality, avoiding hallucinations, and personalization (Figure 3). Although the difference in performance on other perceived code quality metrics was insignificant, we demonstrate that PHIA is substantially less likely to generate code that raises an error. In subsequent sections, we show that PHIA can recover to produce a coherent response even after throwing these errors (Table 2). This shows that PHIA can produce higher quality data analysis code than its base LLM can alone - a finding which underscores the importance of the agent framework for Personal Health Insights.



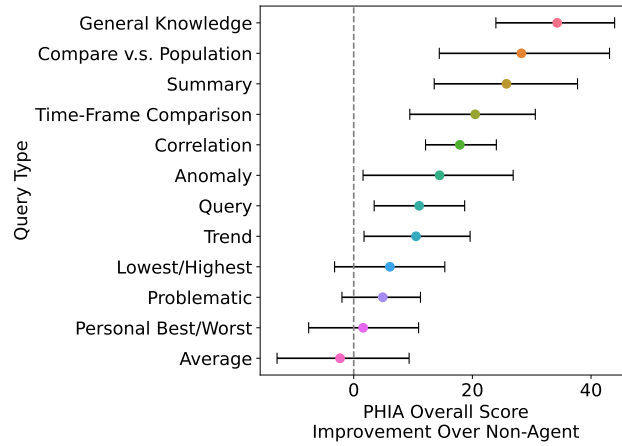


Figure 4: PHIA’s improvement over the non-agent baseline on “Overall Reasoning” for each query type.

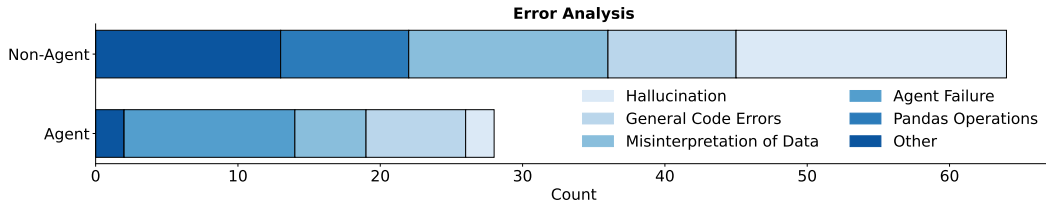


Figure 5: PHIA makes substantially fewer errors than the Non-Agent as determined by expert annotators (Section 2.2)

Model	Error Rate	Recovery Rate
PHIA	0.192 [0.156, 0.228]	0.114 [0.069, 0.161]
Non-Agent All Examples	0.395 [0.362, 0.427]	0.000 [0.000, 0.000]

Table 2: Error Rate (defined as fraction of responses that include at least code one error) is much higher in the Non-Agent model. Furthermore, the agent approach was able to occasionally recover and ultimately provide an answer after code errors, contributing to increased rater perceptions of quality.

To better understand where PHIA’s increased performance comes from we calculated the average “Overall” score from the reasoning evaluation for agent and non-agent methods. We took the difference between these values for each query type (Figure 4). We found that PHIA scores the highest relative to the non-agent baseline on “General Knowledge” and “Compare v.s. Population” questions. Higher “General Knowledge” performance is likely attributable to PHIA’s ability to query web search for outside information and its ability to incorporate parametric knowledge through *Thought* steps. Similarly, the web search tool allows PHIA to search the web for baseline behaviors, like average sleep duration or basal metabolic rate, and incorporate this guidance into “Compare v.s. Population” responses.

**Automatic Analysis of Fatal Code Errors.** We find that PHIA’s error rate is half that of the non-agent baseline (0.192 vs 0.395, Table 2)) indicating that PHIA is much better at generating valid code than the non-agent baseline. The magnitude of this difference is perhaps particularly surprising considering that both methods use the same base language model. This implies that the ability of PHIA to interactively reason about its outputs through “Thought” steps minimizes error-prone code generation.

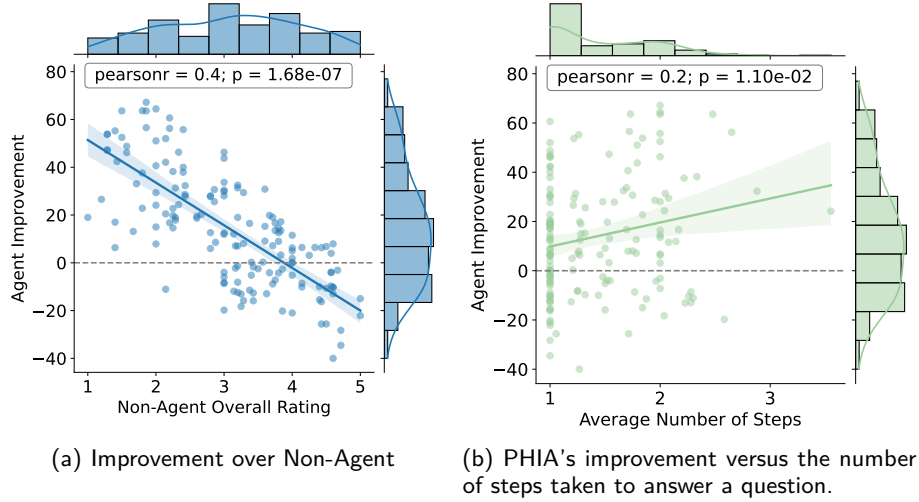


Figure 6: PHIA improves upon questions that the Non-Agent baseline struggles with (Subfigure 6(a)). In particular, PHIA improves most when it uses more reasoning steps to employ additional complexity (Subfigure 6(b)). “Agent Improvement” is the difference in the average “Overall” score for a question between PHIA and the Non-Agent.

One notable advantage of using an agent framework is that PHIA can occasionally recover after it throws a fatal error by interpreting its mistake and correcting it in a subsequent step. To measure this behavior we calculate a “recovery rate” which is number of times that an agent throws a fatal error but ultimately returns an answer divided by the total number of times that a fatal error is thrown. For the Non-Agent this number is zero by construction (since the Non-Agent has only one attempt to respond to the query) but due to its multi-step reasoning ability PHIA’s recovery rate is 0.114 (Table 2). This means that agents like PHIA are more stable with respect to fatal code errors.

**Manual Code Error Analysis.** To determine the nature of these mistakes, two experts independently performed open coding evaluation on all responses from PHIA and the non-agent baseline and were instructed to look for errors, including hallucinations, Python code errors, and misinterpretation of the user’s query. The results were aggregated into one of the following semantic categories: Hallucination, General Code Errors, Misinterpretation of Data, Agent Failure, Pandas Operations, Other. Our results in Table 2 show that PHIA makes less than half the number of errors that the non-agent makes. In particular, PHIA is much less likely to make errors on complex Pandas operations like 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.

## 5 Discussion

Our results suggest that the tool-using and planning abilities of LLM agents like PHIA are effective for analyzing and interpreting personal health data. We observe strong performance on objective personal health insights questions, with PHIA surpassing a Non-Agent baseline’s accuracy by more than XX percentage points. This demonstrates that despite being designed for more complicated tasks multi-step planning ability is useful for addressing even simple queries that require only a single line of code.

This performance improvement extends to complicated open-ended queries. By employing domain experts in wearable data and its analysis we show that PHIA exhibits superior reasoning and analysis code generation properties. This is all the more impressive given that PHIA and the Non-Agent baseline are powered by the same base LLM: Gemini-Advanced. PHIA requires

no additional supervision, only advanced planning abilities and the option to iteratively interact with tools like the Python interpreter and Google search. Therefore as LLMs continue to improve these benefits can be trivially transferred to systems like PHIA.

PHIA is also half as likely as the Non-Agent baseline to produce a fatal code error and is able to recover after making these errors. This elevates the usability of the system and contributes to a higher "Overall" code quality code according to expert annotators.

## 6 Limitations and Future Work

While our works shows 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), but nonetheless we make no claim as to the effectiveness of these insights for modifying behavior or improving health outcomes. Our aim in this paper is to define methods, tasks, and evaluations for agents in personal health. We leave it to future work to evaluate the efficacy of agent methods through a clinical trial.

Furthermore, although our annotators had significant familiarity with the Fitbit ecosystem and Python data analysis, we did not employ exercise or sleep experts to assess the domain-specific validity of PHIA's recommendations. However, the majority of questions in our automatic (Section 2.1) and human (Section 2.2) evaluation datasets are answered through assessment of user data and do not require clinical knowledge. Nonetheless, we acknowledge that before PHIA or a similar agent is deployed as a service for real users care should be taken to verify the accuracy of suggestions where applicable.

In this paper we focus on the analysis of data from wearable devices with Python and explore how that data can be augmented with outside information from Google searches. 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 like smart scales or blood sugar monitors, and more.

## 7 Conclusion

Here we demonstrated the feasibility of PHIA- the Personal Health Insights Agent. We introduced the framework as the first LLM-powered agent that uses code generation and search tools to respond to questions about personal health. We built and publicly released a large dataset of over 4,000 personal health insights questions, and generated privacy sensitive, synthetic wearable data personas that capture the diversity of Fitbit users. We then used expert annotators to show that agent-based methods like PHIA produce responses to personal health questions that are of substantially higher quality than those of their non-agent LLM counterparts. Finally, we discussed the implications of our findings and proposed that agent-based methods like PHIA answer a wide breadth of questions, demonstrate sophisticated reasoning abilities, and are capable of the advanced personalization required to support personal health.

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 [6], and physical inactivity is linked to 9% of premature mortality worldwide [10]. PHIA aims to address these behaviors by bringing the power of language models into everyday lives and empowering users to draw conclusions from their own data.

However, it should not be lost that PHIA is a starting point. As LLMs continue to improve in reasoning ability 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 the first step towards this end.



Figure 7: PHIA is capable of personalizing answers to open-ended health insights questions based on user information. We show two distinct examples of our synthetic wearable users (*Inactive Insomniacs* and *Active Achievers - Low Zzz*) being leveraged as a part of PHIA's reasoning process or final answer.

## References

- [1] T. Althoff, R. Sosič, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec. Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336–339, 2017.
- [2] E. Anderson and J. L. Durstine. Physical activity, exercise, and chronic diseases: A brief review. *Sports medicine and health science*, 1(1):3–10, 2019.
- [3] E. M. BENNETT, R. ALPERT, and A. C. GOLDSTEIN. Communications Through Limited-Response Questioning\*. *Public Opinion Quarterly*, 18(3):303–308, 01 1954.
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [6] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal. The global problem of insufficient sleep and its serious public health implications. *Healthcare*, 7, 2018.
- [7] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023.
- [8] I. R. Galatzer-Levy, D. McDuff, V. Natarajan, A. Karthikesalingam, and M. Malgaroli. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*, 2023.
- [9] M. Guevara, S. Chen, S. Thomas, T. L. Chaunzwa, I. Franco, B. H. Kann, S. Moningi, J. M. Qian, M. Goldstein, S. Harper, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6, 2024.
- [10] I.-M. Lee, E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, and P. 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.
- [11] S. Lee, W. J. Kim, J. Chang, and J. C. Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.
- [12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [13] J. Liu, C. S. Xia, Y. Wang, and L. 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.
- [14] D. McDuff, M. Schaeckermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.
- [15] G. Medic, M. Wille, and M. E. Hemels. Short-and long-term health consequences of sleep disruption. *Nature and science of sleep*, pages 151–161, 2017.
- [16] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.
- [17] Y. Qin, S. Hu, Y. Lin, W. Chen, N. Ding, G. Cui, Z. Zeng, Y. Huang, C. Xiao, C. Han, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.
- [18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. 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.
- [19] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

- [20] W. Shi, R. Xu, Y. Zhuang, Y. Yu, J. Zhang, H. Wu, Y. Zhu, J. Ho, C. Yang, and M. D. Wang. Ehragent: Code empowers large language models for complex tabular reasoning on electronic health records. *arXiv preprint arXiv:2401.07128*, 2024.
- [21] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.
- [22] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.
- [23] E. Stamatakis, M. N. Ahmadi, J. M. Gill, C. Thøgersen-Ntoumani, M. J. Gibala, A. Doherty, and M. Hamer. Association of wearable device-measured vigorous intermittent lifestyle physical activity with mortality. *Nature Medicine*, 28(12):2521–2529, 2022.
- [24] T. Strain, K. Wijndaele, P. C. Dempsey, S. J. Sharp, M. Pearce, J. Jeon, T. Lindsay, N. Wareham, and S. Brage. Wearable-device-measured physical activity and future health risk. *Nature medicine*, 26(9):1385–1391, 2020.
- [25] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.
- [26] M. Swan, T. Kido, E. Roland, and R. P. d. Santos. Math agents: Computational infrastructure, mathematical embedding, and genomics. *arXiv preprint arXiv:2307.02502*, 2023.
- [27] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.
- [28] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2023.
- [29] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.
- [30] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.
- [31] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.
- [32] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models, 2023.
- [33] K. Zhang, N. Patki, and K. Veeramachaneni. Sequential models in the synthetic data vault. *arXiv preprint arXiv:2207.14406*, 2022.
- [34] Y. Zhuang, Y. Yu, K. Wang, H. Sun, and C. Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36, 2024.

## A Overview of Appendices

- Appendix B - **Real-world Data From a Wearable Device.** An example of real-world, daily data from a wearable device. A high-volume, anonymized production form of this kind of data is sampled as a part of our synthetic data creation process (Section 2.3).
- Appendix C - **Examples of Few-shots.** A set of representative few-shot examples that were provided to the methods described in this paper. Between the non-agent and agent (PHIA) examples, the questions are identical.
- Appendix D - **Examples of Agent Behavior.** Additional examples of the agent's (PHIA's) behavior in contrast to non-agent behavior.
- Appendix E - **Raw Data Translations.** The translated raw data viewed by annotators for the evaluation of reasoning quality (??).
- Appendix F - **Annotator Rubrics.** The instructions that were provided to annotators in the reasoning (??) and code (??) evaluations.
- Appendix G - **Inter-Rater Agreement.** Details on our analysis of inter-rater agreement for both human evaluations.
- Appendix H - **Data Schema.** The schema corresponding to the the tables generated as a part of our personas (Section 2.3).

## B Example of Wearable Device Data

As mentioned in Section 2.3, the sequential multivariate and multi-sequence data used to generate synthetic personas for wearable data incorporates stable contextual attributes such as age and gender. An example of such real-world, daily wearable data is shown in Table 3. We sample from high-volume, anonymized production data such as these to ensure a rich representation of real-world variation in our synthetic data.

Pseudo ID	Date	Age	Gender	Step Count	...	Resting Heart Rate
ID-001	2023-10-01	39	Male	5200		75
ID-001	2023-10-02	39	Male	9100	...	74
ID-001	2023-10-03	39	Male	NA	...	NA
...	...	...	...	...	...	...
ID-100	2023-10-01	55	Female	NA		61
ID-100	2023-10-02	55	Female	7200	...	61
ID-100	2023-10-03	55	Female	8600	...	60
...	...	...	...	...	...	...

Table 3: **Real-world Data From a Wearable Device.** An example of real-world, daily data from a wearable device that contains multi-entity, multi-variate mixed data type values with contextual columns and date indices. The data has potential inter-row dependencies and can have missing values. [Tim: anything else we can add that demonstrates complexity well, ideally sth that is then used in one of our examples?]

## C Examples of Few Shots

As described in Section 3, we provided both non-agent and agent models with few-shot examples as a form of guidance toward desirable performance in task execution. In Table 4 below, we provide five examples of both non-agent and agent few-shot examples with varying degrees of tool utility.

---

### Question

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

### Non-Agent Example

**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.

### Agent Example

**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.

---

### Question

*How can I get stronger without gaining mass?*

### Non-Agent Example

**Finish:** A great way to get stronger without gaining mass is to focus on lifting heavy weights with low rep counts. It can also be useful to focus on compound movements instead of isolation exercises.

### Agent Example

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

**Act:** Search('How can I get stronger without gaining mass')

**Observe:** "Below, we are breaking down how to train so that you'll get stronger without gaining muscle mass.

- 1) Focus on Primarily Neuromuscular Adaptions.
- 2) Use load, intensity, sets, and reps, and volume to optimize structural changes.
- 3) Take advantage of metabolic change.

How to build strength without adding bulk. ... Focus on compound lifts. When training for maximum strength, choose compound lifts over isolation exercises. Compound lifts are exercises that use multiple joints and muscle groups simultaneously. ... Use heavy weights in a low rep range for more sets. If you want to become as strong as possible..."

**Finish:** A great way to get stronger without gaining mass is to focus on lifting heavy weights with low rep counts. It can also be useful to focus on compound movements instead of isolation exercises.

---

Continued on next page

---



Table 4 Continued from previous page

<p><b>Question</b>  <i>How does my heart rate effect my health?</i>  <b>Non-Agent Example</b>  <b>Act:</b></p> <pre>Python: request:  -     avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()     avg_resting_heart_rate</pre> <p><b>Observe:</b></p> <pre>71.7741935483871</pre> <p><b>Finish:</b> Your average resting heart rate is approximately 72 beats per minute (bpm), which is within the normal range for adults.  <b>Agent Example</b>  <b>Thought:</b> 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.  <b>Act:</b></p> <pre>Python: request:  -     avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()     avg_resting_heart_rate</pre> <p><b>Observe:</b></p> <pre>71.7741935483871</pre> <p><b>Thought:</b> 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.  <b>Act:</b> Search(request='How does heart rate affect health?')  <b>Observe:</b> "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."  <b>Finish:</b> 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.</p>
---

Continued on next page

Table 4 Continued from previous page

**Question**

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

**Non-Agent Example**

**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.

**Agent Example**

**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.

Continued on next page

Table 4 Continued from previous page

<p><b>Question</b> How is my health overall? <b>Non-Agent Example</b> <b>Act:</b></p>
<pre>Python: request:  -     # Average resting heart rate     avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()     # Average sleep duration and quality     avg_sleep_duration = summary_df['sleep_minutes'].mean()     avg_deep_sleep_percent = summary_df['deep_sleep_percent'].mean()     avg_rem_sleep_percent = summary_df['rem_sleep_percent'].mean()     # Average activity levels     avg_active_zone_minutes = summary_df['active_zone_minutes'].mean()     # Average stress management score     avg_stress_management_score = summary_df['stress_management_score'].mean()     (avg_resting_heart_rate, avg_sleep_duration, avg_deep_sleep_percent,      avg_rem_sleep_percent, avg_active_zone_minutes, avg_stress_management_score)</pre>
<p><b>Observe:</b></p> <pre>(71.7741935483871, 377.741935483871, 22.846458435321964, 21.5051126153761, 73.0,  5.67741935483871)</pre>
<p><b>Finish:</b> Your overall health appears to be good. Your average resting heart rate of 71.77 bpm is within a healthy range. You're getting an average of 377.74 minutes of sleep per night, with 22.85% in deep sleep and 21.51% in REM sleep, indicating adequate sleep quality. Your average active zone minutes are 73.0, suggesting you engage in regular physical activity. Your average stress management score of 5.68 indicates you're managing stress effectively. It's important to maintain these healthy habits and consult with a healthcare professional for any specific concerns or changes in your health status.</p>
<p><b>Agent Example</b> <b>Thought:</b> This question requires a comprehensive analysis of various health metrics available in the summary_df DataFrame, including resting heart rate, sleep quality, activity levels, and stress management score. Once I have that information I will use the search tool to pull in additional context. <b>Act:</b></p>
<pre>Python: request:  -     # Average resting heart rate     avg_resting_heart_rate = summary_df['resting_heart_rate'].mean()     # Average sleep duration and quality     avg_sleep_duration = summary_df['sleep_minutes'].mean()     avg_deep_sleep_percent = summary_df['deep_sleep_percent'].mean()     avg_rem_sleep_percent = summary_df['rem_sleep_percent'].mean()     # Average activity levels     avg_active_zone_minutes = summary_df['active_zone_minutes'].mean()     # Average stress management score     avg_stress_management_score = summary_df['stress_management_score'].mean()     (avg_resting_heart_rate, avg_sleep_duration, avg_deep_sleep_percent,      avg_rem_sleep_percent, avg_active_zone_minutes, avg_stress_management_score)</pre>
<p><b>Observe:</b></p> <pre>(71.7741935483871, 377.741935483871, 22.846458435321964, 21.5051126153761, 73.0,  5.67741935483871)</pre>
<p><b>Thought:</b> I should see if this is enough REM Sleep <b>Act:</b> Search(How much REM sleep should I be getting?) <b>Observe:</b> "What is REM Sleep, Why it's Important, How Much You Need On average you'll go through 3-5 REM cycles per night, with each episode getting longer as the night progresses. The final one may last roughly an hour. For healthy adults, spending 20-25% of your time asleep in the REM stage is a good goal. If you get 7-8 hours of sleep, around 90 minutes of that should be REM. How much REM sleep do you need? There's no official agreement on how much REM sleep you need, but this stage is important for dreaming, and experts believe dreaming helps you process emotions. For most adults, REM takes up about 25% of sleep, which seems to be healthy during average sleep cycles. Which is better, REM or deep sleep? All the stages of sleep are necessary, and none is better than any other. You need a balance of around 25% REM and 25% of the deepest NREM sleep to maintain your health and wellbeing." <b>Thought:</b> By analyzing these metrics, I can provide an overview of your overall health status. <b>Finish:</b> Your overall health appears to be good. Your average resting heart rate of 71.77 bpm is within a healthy range. You're getting an average of 377.74 minutes of sleep per night, with 22.85% in deep sleep and 21.51% in REM sleep, indicating adequate sleep quality. Your average active zone minutes are 73.0, suggesting you engage in regular physical activity. Your average stress management score of 5.68 indicates you're managing stress effectively. It's important to maintain these healthy habits and consult with a healthcare professional for any specific concerns or changes in your health status.</p>

Table 4: **Examples of Few-shots.** Comparison of few-shot examples between Non-Agent and Agent (PHIA).

## D Examples of Agent Behavior

Table 5 consists of contrastive examples of non-agent and agent (PHIA) behavior when presented with the same question. These are the kinds of responses that were evaluated by human annotators as described in ?? and ??.

### Question

*If I noticed spikes in my data, what action should I take?*

### Non-Agent Example

**Finish:** If you notice spikes in your health data, it's important to consider the following actions:

**Identify the cause:** Try to determine what might have caused the spike, such as a change in activity level, diet, or stress.

**Monitor your symptoms:** Pay attention to any other symptoms you may be experiencing, such as fatigue, pain, or shortness of breath.

**Consult a healthcare professional:** If the spikes are persistent or accompanied by other symptoms, it's essential to seek medical advice. They can help you determine the underlying cause and recommend appropriate treatment options.

**Adjust your lifestyle:** If the spikes are related to lifestyle factors, such as exercise or diet, consider making adjustments to improve your overall health.

### Agent Example

**Thought:** This question cannot be answered using Fitbit data. I will use the Search tool to find general information about spikes in health data.

**Act:** Search(request='What to do if you notice spikes in health data?')

**Observe:** "It's potentially a critical issue, so you may need to begin a parallel process. Respond to the issue as if it was genuine, but also begin the process of investigation. Remember to clearly communicate to others the extent to which you have been able to verify the data. Scenario B: It's significant, but doesn't need urgent attention. If you responded to every blip in your data immediately, you would never get anything else done. But that's okay. Schedule a time to investigate the issue (and make sure you follow through.) Scenario C: There wouldn't be any significant consequences.

Have we changed anything? Start internally. Check for new updates, new features, areas of work. And pay very close attention to dates and sequences of events. This will allow you to rule out working theories if the sequences of events don't match. It will also reveal coincidences, such as two unusual but seemingly unconnected things happening at the same time. Always interrogate these coincidences. It's important to keep a completely open mind. The cause of your spike may be an unintended consequence of something completely unrelated. Normally, it's because something, somewhere, has changed.

This level stayed high for 2 days in 96% of the individuals. If you like this content and want more, I am doing a special life-time membership offer. Click here: <https://www.drbeen.com/yt-special-p/> Want to support this work?: Buy me a coffee :) <https://www.buymeacoffee.com/DrMobeenSyed> Become my patron: [https://www.patreon.com/mobeensyed?fan\\_landing=true](https://www.patreon.com/mobeensyed?fan_landing=true) PayPal: <https://www.paypal.com/donate/?business=DrMobeenSyed> Please consult with a physician or healthcare professional regarding any medical or mental health related diagnosis or treatment. No information in this video should ever be considered as a substitute for advice from a healthcare professional. Author's highlights: Vaccination confers broader IgG binding of variant RBDs than SARS-CoV-2 infection Imprinting from initial antigen exposures alters IgG responses to viral variants Histology of mRNA vaccinee lymph nodes shows abundant GCs Vaccine spike antigen and mRNA persist for weeks in lymph node GCs Study highlights for the spike protein in the blood: Important points ... Here we will continue to observe the spike protein's levels in the blood. Figure 7 J, K, and L diagrams are important. URL list from Wednesday, Mar. 16 2022 DrBeen: Continuing Medical Education Online | USMLE Prep | CME and CE marketplace | Nurse Practitioner and Physician Assistant Training Programs | Medical Student Training <https://www.drbeen.com/> Immune imprinting, breadth of variant recognition, and germinal center response in human SARS-CoV-2 infection and vaccination - ScienceDirect <https://www.sciencedirect.com/science/article/pii/S0950268821001111>

**Finish:** If you notice spikes in your health data, it's important to take the following steps:

**Respond as if the issue is genuine:** Take immediate action to address any potential health concerns.

**Investigate the cause:** Look for any changes in your lifestyle, environment, or medications that may have contributed to the spikes.

**Communicate with others:** Inform your healthcare provider or other relevant individuals about the spikes and the extent to which you have been able to verify the data.

**Schedule a follow-up:** If the spikes are significant but don't require urgent attention, schedule a time to investigate the issue further.

**Consider all possibilities:** The cause of the spikes may be unrelated to your health, so keep an open mind and consider all potential factors.

It's also important to note that not all spikes in health data are cause for concern. Some fluctuations are normal and may be due to factors such as stress, exercise, or changes in diet. However, if you notice persistent or significant spikes, it's always best to consult with a healthcare professional.

Continued on next page

Table 5 Continued from previous page

**Question**

*What are my average calories burned per week?*

**Non-Agent Example**

**Act:**

```
Python:
request: |-
activities_df.index = activities_df.index.date
activities_df['week'] = activities_df.index.week
weekly_calories = activities_df.groupby('week')['calories'].sum()
average_weekly_calories = weekly_calories.mean()
average_weekly_calories
```

**Observe:**

```
#ERROR#: AttributeError: 'Index' object has no attribute 'week'
```

**Finish:** I'm sorry, but I'm unable to answer your question because the provided code contains an error. The error message indicates that the 'Index' object does not have a 'week' attribute.

**Agent Example**

**Thought:** I can use the 'during' method for this.

**Act:**

```
Python:
request: |-
# Calculate the number of weeks in the data
num_weeks = (summary_df.index.max() - summary_df.index.min()).days / 7
# Calculate the average calories burned per week
avg_calories_per_week = activities_df["calories"].sum() / num_weeks
avg_calories_per_week
```

**Observe:**

```
744.3333333333334
```

**Finish:** Your average calories burned per week is 744.33.

Continued on next page

Table 5 Continued from previous page

<p><b>Question</b>  <i>How does my Stress Score correlate with my daily Steps?</i></p> <p><b>Non-Agent Example</b>  <b>Act:</b></p> <pre> Python: request:  - summary_df['date'] = summary_df.index.date joined = summary_df.join(activities_df, how='right', lsuffix='_activity') correlation = joined['stress_management_score'].corr(joined['steps']) correlation </pre> <p><b>Observe:</b></p> <pre>nan</pre> <p><b>Finish:</b> There is no data available to calculate the correlation between your Stress Score and daily Steps.</p> <p><b>Agent Example</b>  <b>Thought:</b> To answer this question, I'll first join the summary_df and activities_df DataFrames. Then, I'll calculate the correlation between the stress_management_score and steps columns.  <b>Act:</b></p> <pre> Python: request:  - activities_df.index = activities_df.index.date joined = summary_df.join(activities_df, how='right', lsuffix='_activity') correlation = joined['stress_management_score'].corr(joined['steps']) correlation </pre> <p><b>Observe:</b></p> <pre>-0.37575756401643967</pre> <p><b>Thought:</b> The correlation coefficient indicates a moderate negative correlation between stress score and daily steps.  <b>Finish:</b> There is a moderate negative correlation of -0.38 between your stress management score and daily steps. This suggests that as your daily step count increases, your stress management score tends to decrease. In other words, taking more steps may help you manage stress better.</p>
--

Continued on next page

Table 5 Continued from previous page

<div><div>Question</div><div>Does doing abs everyday help my health goal?</div><div>Non-Agent Example</div><div>Act:<div><div>Python:</div><div>request:  - print("I'm sorry, but I cannot provide medical advice. Please consult a healthcare professional for guidance on health goals.")</div></div><div>Observe:<div>#ERROR#: Compilation failure</div></div><div>Finish: I'm sorry, but I cannot provide medical advice. Please consult a healthcare professional for guidance on health goals.</div><div>Agent Example</div><div>Thought: This question cannot be answered using Fitbit data. I will use the Search tool to find relevant information.</div><div>Act: Search(request="Does doing abs everyday help my health goal?")</div><div>Observe: "The Power of Consistency: Benefits of Daily Abdominal</div><div>Incorporating daily abdominal workouts into your fitness routine is a commitment that can yield remarkable benefits. From enhanced core strength and better posture to reduced back pain and improved athletic performance, the advantages are clear.</div><div>It sounds like a late-night commercial for a miracle product, but strong abs may actually help you live longer. Researchers at Mayo Clinic have noted that those with larger waist circumference size tended to have a greater risk of premature death than those with smaller waists. This was true even in the face of other risk factors like smoking, being sedentary, and alcohol use. Could ab workouts now keep you around in the long term? It's worth a shot."</div><div>Finish: Doing abdominal exercises daily can have several health benefits, including:</div><div><div>Enhanced core strength</div><div>Better posture</div><div>Reduced back pain</div><div>Improved athletic performance</div></div><div>Additionally, research suggests that strong abdominal muscles may be associated with a lower risk of premature death. However, it's important to note that abdominal exercises alone are not a comprehensive solution for overall health and should be combined with a balanced diet and regular physical activity.</div></div></div>
--

Continued on next page

Table 5 Continued from previous page

**Question**

*I am noticing my HRV (while asleep) seems to be all over the place and I am not sure what's affecting it and what it means?*

**Non-Agent Example**

**Act:**

```
Python:
request: |-
    # Calculate the standard deviation of HRV
    hrv_std = summary_df['heart_rate_variability'].std()
    # Find days with HRV more than 1 standard deviation from the mean
    high_hrv_days = summary_df[summary_df['heart_rate_variability'] > (summary_df['heart_rate_variability'].mean() + hrv_std)]
    low_hrv_days = summary_df[summary_df['heart_rate_variability'] < (summary_df['heart_rate_variability'].mean() - hrv_std)]
    # Print the dates with high and low HRV
    print("Days with high HRV:", high_hrv_days.index.date)
    print("Days with low HRV:", low_hrv_days.index.date)
```

**Observe:**

Days with low HRV:

**Finish:** Your HRV (heart rate variability) is a measure of the variation in time between heartbeats. A higher HRV is generally considered a sign of good health, while a lower HRV may indicate stress or other health issues. The following dates had unusually high HRV: The following dates had unusually low HRV: Factors that can affect HRV include:

Stress

Sleep quality

Exercise

**Agent Example**

**Thought:** To understand the variability in the user's HRV during sleep, I'll first analyze the data to identify any patterns or trends. Then, I'll use the Search tool to gather information about factors that can affect HRV.

**Act:**

```
Python:
request: |-
    # Calculate the standard deviation of HRV during sleep
    sleep_hrv = summary_df["heart_rate_variability"][summary_df["sleep_minutes"] > 0]
    sleep_hrv_std = sleep_hrv.std()
    sleep_hrv_std
```

**Observe:**

15.517624542618865

**Thought:** The standard deviation of the user's HRV during sleep is relatively high, indicating significant variability. Now, I'll use the Search tool to find information about factors that can affect HRV.

**Act:** Search(request="What factors affect HRV?")

**Observe:** "Analysis of Heart Rate Variability and Implication of Different

The HRV is influenced by various variables such as; pathological, physiological, psychological, environmental factors, lifestyle factors, and genetic factors, etc.

Factors Influencing Heart Rate Variability. Factors that affect your heart rate variability include: External and lifestyle factors. Several outside and lifestyle factors that can change your heart rate variability. For instance, the following tend to lower your heart rate variability: More social stress, anxiety, and depression. Being in a hot environment (cold environments don't seem to affect it very much) Exposure to noise. Having more body weight or more fat mass. Drinking more than one to two drinks of alcohol per day, depending on your size and weight. Drinking coffee (but it may not be by very much) Smoking. Pain.

Physiological factors. Physiological factors that affect your heart rate variability include age, sex assigned at birth, and circadian rhythm. For instance: Your heart rate variability tends to go up until you reach age 15, after which it goes down. It continues to go down throughout the rest of your life. People assigned female at birth tend to have a slightly lower heart rate variability than those assigned male at birth. Heart rate variability tends to be higher when you sleep than after you wake up."

**Finish:** Your heart rate variability (HRV) during sleep shows significant variation, with a standard deviation of 15.52. HRV can be influenced by various factors, including:

**Lifestyle factors:** Stress, anxiety, alcohol consumption, caffeine intake, smoking, and body weight.

**Physiological factors:** Age, sex, and circadian rhythm.

**Environmental factors:** Heat and noise exposure.

To understand the specific factors affecting your HRV, you may consider tracking your lifestyle habits and environmental conditions alongside your HRV measurements. Additionally, consulting with a healthcare professional can help you interpret your HRV data and identify any underlying health concerns.

Table 5: **Examples of Agent Behavior.** Comparison of responses from Non-Agent versus Agent (PHIA).



## E Raw Data Translation

In Table 6, we present examples of raw data that were translated for human annotators evaluating the reasoning quality of non-agent and agent (PHIA) models (??). 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 stress affect the ability to stay asleep?*

**Raw Data**  
**Act:**

```
Python:
request: |-
    # Calculate the correlation between stress management score and awake minutes
    corr_stress_awake = summary_df['stress_management_score'].corr(summary_df['
    awake_minutes'])
    corr_stress_awake
```

**Observe:**

```
-0.16624525699385023
```

**Finish:** The correlation between stress management score and awake minutes is -0.166, indicating a weak negative correlation. This suggests that as stress levels increase, the number of minutes spent awake during the night may slightly decrease. However, it's important to note that correlation does not imply causation, and other factors may also influence the ability to stay asleep.

**Translation**  
**Overall Goal:** This code aims to determine the correlation between stress and sleep quality.

**Basic Logic:** The code calculates the correlation coefficient between a "stress management score" and the number of "awake minutes" recorded. A correlation coefficient measures the strength and direction of the relationship between two variables.

**Final Deliverable:** The code outputs a correlation coefficient of -0.166, indicating a weak negative correlation. This suggests that higher stress levels may be slightly associated with fewer minutes spent awake during the night. However, it's crucial to note that this is just a correlation and does not prove a causal relationship between stress and sleep quality.

**Question**  
*I tend to be drinking more water than usual. Has this led to better health and fitness?*

**Raw Data**  
**Act:**

```
Python:
request: |-
    water_df = activities_df[activities_df["activityName"] == "Water"]
    water_df["water_amount"] = water_df["calories"]
    # Calculate the correlation between water intake and resting heart rate
    corr_water_resting_heart_rate = water_df["water_amount"].corr(summary_df["
    resting_heart_rate"])
    # Calculate the correlation between water intake and active zone minutes
    corr_water_active_zone_minutes = water_df["water_amount"].corr(summary_df["
    active_zone_minutes"])
    (corr_water_resting_heart_rate, corr_water_active_zone_minutes)
```

**Observe:**

```
(nan, nan)
```

**Finish:** Unfortunately, the provided data does not contain information about water intake, so I cannot determine if increased water consumption has led to better health and fitness.

**Translation**  
**Overall Goal:** This code aims to determine if increased water intake has led to improvements in health and fitness.

**Basic Logic:** The code tries to find data on water intake and compare it to data on health metrics like resting heart rate and active zone minutes. If the data is available, it calculates the correlation between water intake and these metrics.

**Final Deliverable:** The code checks if there is data on water intake. If there is, it calculates the correlation between water intake and health metrics. Otherwise, it states that the data does not contain information on water intake and cannot answer the question.

Continued on next page

Table 6 Continued from previous page

**Question**

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

**Raw Data**

**Thought:** To find the personal bests, I can use the 'max()' function on the relevant columns in the activities\_df DataFrame.  
**Act:**

```
Python:
request: |-
    # 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)
```

**Finish:** 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.

**Question**

*Why am I not losing weight?*

**Raw Data**

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

**Act:** Search('Why am I not losing weight')

**Observe:** '13 Reasons You're 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. You're not eating whole foods. Food quality is just as important as quantity. ... Keep in mind that many processed foods labeled as "health foods" aren't 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.

**Finish:** 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.

If you are struggling to lose weight, it's important to consult with a healthcare professional to determine the underlying cause and develop an appropriate plan.

**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.

Continued on next page

Table 6 Continued from previous page

<p><b>Question</b> How can I improve my sleep times (i.e., waking up and sleeping earlier)?</p> <p><b>Raw Data</b> <b>Thought:</b> To improve sleep times, I'll first search for general recommendations on improving sleep habits. Then, I'll analyze the user's sleep data from the summary_df DataFrame to identify any specific patterns or areas for improvement. <b>Act:</b> Search(request='How to improve sleep times') <b>Observe:</b> ""Some habits that can improve your sleep health:</p> <p>Be consistent.</p> <p>Make sure your bedroom is quiet, dark, relaxing, and at a comfortable temperature.</p> <p>Remove electronic devices, such as TVs, computers, and smart phones, from the bedroom.</p> <p>Avoid large meals, caffeine, and alcohol before bedtime.</p> <p>Get some exercise. Tips for Better Sleep   CDC. Good sleep habits (sometimes referred to as "sleep hygiene") can help you get a good night's sleep. Some habits that can improve your sleep health: Be consistent. Go to bed at the same time each night and get up at the same time each morning, including on the weekends. Make sure your bedroom is quiet, dark, relaxing, and at a comfortable temperature. Remove electronic devices, such as TVs, computers, and smart phones, from the bedroom. Avoid large meals, caffeine, and alcohol before bedtime. Get some exercise. Being physically active during the day can help you fall asleep more easily at night."</p> <p><b>Thought:</b> The search results provide general recommendations for improving sleep habits, such as maintaining a consistent sleep schedule, creating a relaxing sleep environment, and avoiding caffeine and alcohol before bed. Now, let's analyze the user's sleep data to identify any specific patterns or areas for improvement. <b>Act:</b></p> <div> <pre>Python: request:  -     # Calculate the average bed time and wake-up time     avg_bed_time = summary_df['bed_time'].mean()     avg_wake_up_time = summary_df['wake_up_time'].mean()     # Convert the average times to a 24-hour format     avg_bed_time_24h = avg_bed_time.strftime('%H:%M')     avg_wake_up_time_24h = avg_wake_up_time.strftime('%H:%M')     (avg_bed_time_24h, avg_wake_up_time_24h)</pre> </div> <p><b>Observe:</b></p> <div> <pre>('22:00', '05:35')</pre> </div> <p><b>Finish:</b> Based on the general recommendations and your specific sleep data, here are some tips to improve your sleep times:</p> <p><b>Maintain a consistent sleep schedule:</b> Your average bed time is 10:00 PM and your average wake-up time is 5:35 AM. Try to stick to this schedule as much as possible, even on weekends.</p> <p><b>Create a relaxing sleep environment:</b> Make sure your bedroom is dark, quiet, and cool. Consider using blackout curtains, a white noise machine, or earplugs to minimize distractions.</p> <p><b>Avoid caffeine and alcohol before bed:</b> Caffeine and alcohol can interfere with sleep. Avoid consuming these substances in the hours leading up to bedtime.</p> <p><b>Get regular exercise:</b> Physical activity can help you fall asleep more easily at night. However, avoid exercising too close to bedtime.</p> <p>If you consistently have trouble sleeping, consider consulting with a healthcare professional. There may be underlying medical conditions or other factors that are affecting your sleep.</p> <p><b>Translation</b> <b>Overall Goal:</b> The code aims to provide suggestions on how to improve sleep times by waking up and going to bed earlier.</p> <p><b>Basic Logic:</b> The code first searches for general recommendations on improving sleep habits. Then, it analyzes the user's sleep data to calculate the average bed and wake-up times.</p> <p><b>Final Deliverable:</b> Based on the search results and sleep data analysis, the code provides personalized tips for improving sleep times, such as maintaining a consistent sleep schedule, creating a relaxing sleep environment, avoiding caffeine and alcohol before bed, and getting regular exercise. It also advises consulting a healthcare professional if sleep problems persist.</p>
---

Table 6: **Raw Data Translation.** Examples of raw data translated into a form that is interpretable by human annotators.

## F Annotator Rubrics

Table 7 and Table 8 detail the rubrics given to annotators in order to evaluate the performance of both a non-agent and an agent (PHIA) model in terms of reasoning quality (?? and code quality (??) respectively. Table 7 focuses on evaluating the relevance, domain knowledge, harmfulness, and clarity of textual responses. In contrast, Table 8 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	What is the overall quality of this response? Options: 1-5	<ol style="list-style-type: none"> <li>1. Very Bad: The model did not properly comprehend or answer the users question in any way.</li> <li>2. Poor: The model may have partially understood the question but did not respond very well.</li> <li>3. Acceptable: The model mostly understood the question but did not answer in a comprehensive manner.</li> <li>4. Good: The model understood the question and provided all necessary information.</li> <li>5. Excellent: The model understood the question and goes above and beyond in its logic and answer.</li> </ol>
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	Accounts for correct logic and missing logic
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 7: **Reasoning Quality Rubric.** Questions used for annotating the reasoning quality (??) of final answers.

## G Inter-Rater Agreement

In order to gauge the reliability of the ratings provided, we used Bennett's S-Score [3] which is especially useful to assess how consistent individuals are in making categorical judgments.

Dimension	Question and Options	Comments
Overall	What is the overall quality of the code in this response? Options: 1-5	<ol style="list-style-type: none"> <li>1. Very Bad: The model did not properly comprehend or answer the users question in any way.</li> <li>2. Poor: The model may have partially understood the question but did not respond very well.</li> <li>3. Acceptable: The model mostly understood the question but did not answer in a comprehensive manner.</li> <li>4. Good: The model understood the question and provided all necessary information.</li> <li>5. Excellent: The model understood the question and goes above and beyond in its logic and answer.</li> </ol>
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 <code>heart_rate_variability</code> 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 8: **Code Quality Rubric.** Questions used for annotating the code quality (??) of final answers.

Bennett's S-Score [3] 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 [3] 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 63% better than what would be expected by random chance alone and is a considerably good degree of agreement. Table 9 summarizes the inter-rater agreement using Bennett's S-Score for six categories of human ratings on query responses and five categories of human ratings on code generations.

## H Data Schema

Table 10 and Table 11 correspond to descriptions of daily summary data and activities data respectively. This is structured data that both non-agent and agent (PHIA) models view and

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

Table 9: **Inter-Rater Agreement.** Bennett's S Scores [3] for human ratings of the query responses and code generations.

processes as a part of their workflow. In Table 10, 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 11 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" heart 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 10: **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 11: **Activities Table Schema.** Columns, data types and data descriptions in the Activities table.