Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications

Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, Longfei Shangguan
University of Pittsburgh, University of Illinois Urbana-Champaign, Independent Researcher, Google

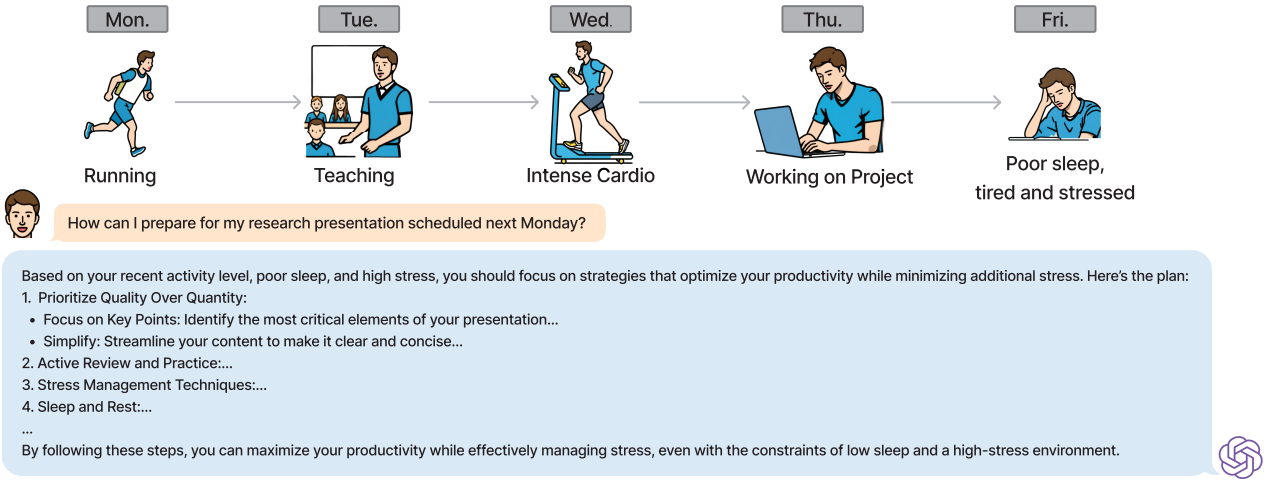

Figure 1: A PhD student's weekly schedule. On Monday, Bob ran 2 miles after a whole day of literature survey. On Tuesday, feeling a little stressed after attending an intensive workshop and lab session. On Wednesday, Bob joined an intense cardio workout to get sort of relax after attending two classes. On Thursday, he was busy with his project and had no time for exercise. By Friday, he had a poor sleep, feeling tired and stressed due to an upcoming presentation. So, he decided to ask WellMax's for advice on preparing his research presentation scheduled for next Monday. The icons are generated by Stable Assistant[4].

ABSTRACT

This paper explores sensor-informed personal agents that can take advantage of sensor hints on wearables to enhance the personal agent's response. We demonstrate that such a sensor-in-the-loop AI agent design can be easily integrated into existing LLM agents by building a prototype named WellMax based on existing well-developed techniques such as structured prompt templates and few-shot prompting. The head-to-head comparison with a non-sensor-informed agent across five use scenarios demonstrates that this sensor-in-the-loop design can effectively improve users' needs and their overall experience. The deep-dive into agents' replies and participants' feedback further reveals that sensor-in-the-loop agents not only provide more contextually relevant responses but also exhibit a better understanding of user priorities and situational nuances. In addition, we conduct two case studies to examine the potential pitfalls and distill key insights from this sensor-in-the-loop agent. We hope this work can spawn new ideas for building more intelligent, empathetic, and effective AI-driven personal assistants.

CCS CONCEPTS

• **Human-centered computing** → **Personal digital assistants**.

KEYWORDS

Context-aware LLMs, Personalized AI assistants, Wearable sensing, Mobile and ubiquitous computing

ACM Reference Format:
Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, Longfei Shangguan. 2025. Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715014.3722082

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/2025/05.
https://doi.org/10.1145/3715014.3722082

1 INTRODUCTION

Large Language Models (LLMs) have remarkably boosted the personal agent's capability on answering questions [24, 44]. Recently, with the advent of miniaturized LLMs, such as Google's Gemma-2 [40] and Apple's OpenELM [36], coupled with the growing computational power of System-on-Chips (SoCs), the deployment of LLMs on mobile devices has become increasingly feasible. We envision a future world where every mobile device will host an LLM-powered personal agent tailored to its user. Like Javis to Iron Man, we believe this personal agent will serve as the user's digital twin, sensing and interpreting the user's rich context (e.g., physiological signals and physical activities) around the clock and providing just-in-time assistance when needed.

While the early signs are positive and the industrial wheels are in motion, current LLM-powered personal agents still struggle to correctly comprehend users' intentions and questions as our human beings are, particularly when queries are vague or ambiguous [10,

25, 54]. For example, in conversations, we often ask questions like *"Why isn't this working?"* when we fail to connect phones to Wi-Fi, expecting an answer from friends nearby. While our friends can understand this concise yet vague question since they know the context through visual and auditory cues, personal agents usually fail to do so as they lack the ability to sense physical context when answering questions. As a result, often the time we have to change our question-asking behaviors by carefully rephrasing our inquiries, e.g., rephrasing *"Why isn't this working?"* as *"Why won't my phone connect to Wi-Fi even though the Wi-Fi signal is very strong?"*

These and many similar cases lead us to think: *Is it possible for mobile users to converse with their agents naturally, as if they were talking to their friends?* While Prompt Engineering [28, 45, 46] and Retrieval-Augmented Generation (RAG) [16] are crucial techniques for achieving this goal, strategies that compensate for these ***user physical context-awareness*** difficulties are often overlooked. For instance, the question *"Why isn't this working?"* could mean drastically different things depending on the user's physical and physiological activities, while current prompt engineering and RAG solutions would not correctly interpret it if they have no clue about the user's current physical context.

Our key idea is that despite extracting relevant information from users' query history and external documents, personal agents can use sensor hints to get another degree of user context and optimize their responses accordingly. We believe this *sensor-in-the-loop* approach is practical and readily implementable because almost every smartphone, wearable, and tablet today is equipped with a wide spectrum of sensors like GPS, accelerometers, gyroscopes, microphones, and so on. These sensors are used by various mobile applications to track user physical/physiological activities, providing crucial physical and physiological user context.

While the high-level idea of this sensor-in-the-loop LLM agent is intuitive, many open questions remain unanswered. For instance, can the current LLM models effectively interpret these sensory cues to deliver personalized responses? Are techniques like in-context learning or parameter-efficient fine-tuning, *etc.*, adequate with minor modification? Or is a clean-slate design necessary? Would model size matter? In what situations would these techniques be most effective? Additionally, how do mobile users perceive these sensor-enhanced responses? Would they prefer them over conventional LLM agent's responses all the time? What are the potential challenges in integrating sensory data into the LLM's decision-making process?

In this paper, we conduct a pioneer study on this sensor-in-the-loop LLM agent based on an open wearable dataset LifeSnaps [52]. Our first goal is to de-risk the core capability – suppose the LLM agent could obtain accurate contextual information about the mobile user derived from his or her sensory data, is it possible to provide context-aware and personalized response? To this end, we develop a prototype called WellMax that leverages in-context learning to rewrite the user's query/question by incorporating the physical and physiological contexts. Our findings demonstrate that with simple regularized few-shot prompting and prompt rewriting, the current LLM-powered personal agents such as GPT-4o can effectively adapt their behavior and responses to sensor-derived hints, resulting in more contextually aware, empathetic, and responsive replies. Figure 1 shows a future AI agent application scenario we aim to create, out of many.

Our second goal is to answer the aforementioned open questions by thoroughly analyzing these personalized responses in different use cases. To achieve this, we select five distinct user scenarios that represent a diverse range of everyday activities that require different types of contextual understanding and personalized assistance: Exercise Routine Planning, Healthy Eating Planning, Presentation Preparation, After-Work Planning, and Book Recommendation. Each scenario highlights a unique combination of user needs and contextual variables, such as physical activity levels, stress management, and sleep quality. These scenarios provide a broad spectrum for evaluating the effectiveness of the sensor-in-the-loop approach. We then design a questionnaire based on both WellMax's and baseline agent's responses and recruit participants from UserTesting platform [42] *under the approval of our Internal Review Board (IRB)* to assess the clarity, context-awareness, empathy, trustworthiness, and overall satisfaction of the responses from two agents.

Experiment results reveal that WellMax achieves higher overall satisfaction scores across GPT-4o powered agents, particularly in context awareness and empathy. Specifically, the mean overall satisfaction rating across all 5 scenarios for WellMax with GPT-4o is 6.0 ± 1.3 (out of 7), compared to 5.2 ± 1.6 for the baseline approach. A Wilcoxon signed-rank test confirms this difference is significant ($p < 0.05$), and the corresponding Cohen's d indicates a medium effect size. These results highlight the significant improvements in context-awareness and empathy ratings brought by leveraging a *sensor-in-the-loop* approach in personal assistants. We further conducted two case studies: one examining the *serial-position effect* in response structure, and another exploring WellMax 's performance on smaller language models. These explorations help identify potential pitfalls and refine our understanding of how sensor-driven personalization can optimize user interactions.

It is worth noting that we are not the first to explore LLM's comprehension capability on sensing data. Previous pioneering works have already utilized LLMs to interpret IoT and wearable sensing data, enabling reasoning about the physical world [48] and human activities [19, 21, 23, 32, 38], and the development of domain-specific health agents [14, 22, 26, 37, 47, 50]. In contrast, we focus on integrating mobile users' physical and physiological activity histories derived from sensory data with the reasoning capabilities of LLMs to enhance the user experience in interactions with personal agents. On the other hand, our work does not aim to introduce new designs; instead, it assesses the feasibility of developing a sensor-in-the-loop personal agent using established methods. Our focus is on evaluating the viability of this prototype, gathering user feedback, and identifying potential limitations through extensive benchmarking studies. We hope the results and takeaways drawn from our comprehensive user studies can spawn new ideas for building intelligent, empathetic, and effective personal AI agents.

2 A SENSOR-IN-THE-LOOP LLM AGENT

We demonstrate how this sensor-in-the-loop paradigm can be seamlessly integrated into existing LLM agents by implementing WellMax. WellMax contains two components: (i): **Sensing Results Processing Module**, which categorizes and interprets different

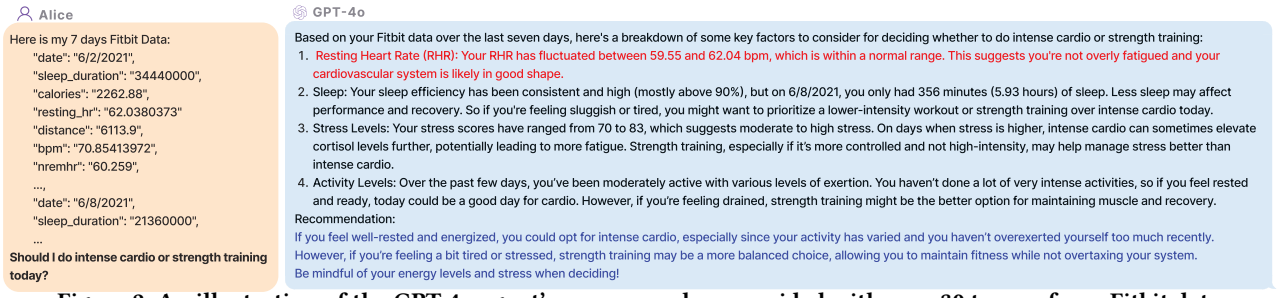

Figure 2: An illustration of the GPT-4o agent's response when provided with over 30 types of raw Fitbit data.

types of sensory information to infer the user's persona; (ii): **Query Rewriting Module**, which modifies user queries based on this inferred persona. We elaborate on each design component below.

2.1 Sensing Results Processing Module

Existing wearable devices such as Apple Watch [6] can track a diverse set of user states, including gym activities, walking distance, and calorie expenditure, while also monitoring vital signs such as heart rate and breath through various onboard sensors. Many prior works [8, 13, 39, 43] have demonstrated these capabilities with high accuracy. Here, we assume these physical and physiological contexts are readily available, and our goal is to build a processing module to preprocess these data before feeding them to the LLM.

Why not feed the user's query along with detailed sensor context to the LLM agent? Given the impressive comprehension capabilities of LLMs across various tasks [2, 11, 41], an intuitive solution is to provide the LLM agent with the entire detailed sensor context, allowing it to identify the most relevant parts for the user's query. However, we find that this approach does not consistently yield accurate results.

As shown in Figure 2, a mobile user asks the LLM agent (GPT-4o) the question: *"Should I do intense cardio or strength training today?"*, along with seven days of Fitbit data. As highlighted in red, the LLM identifies important metrics like Resting Heart Rate (RHR), sleep duration, and stress levels but treats them in isolation. It notes, *"Your RHR has fluctuated between 59.55 and 62.04 bpm, which is within a normal range.* ***This suggests you're not overly fatigued*** *and your cardiovascular system is likely in good shape."* However, this conclusion ignores other critical factors such as high stress levels and inconsistent sleep patterns—today showing only 356 minutes (5.93 hours) of sleep. These factors should be considered together to assess physical readiness, but the LLM fails to combine them into a comprehensive analysis.

This lack of integration leads to an overly conditional and generic recommendation (marked in blue in Figure 2). The LLM suggests that *"if you feel well-rested and energized, you could opt for intense cardio,"* but also offers that *"if you're feeling a bit tired or stressed, strength training may be a more balanced choice."* This vague advice places the decision back on the user rather than providing a clear, data-driven recommendation. Given the user's high stress and poor sleep, the LLM should have prioritized a specific suggestion, like focusing on recovery. The failure to properly integrate the data results in a lack of decisive and personalized guidance.

Instead of feeding all captured sensor context to the LLM agent and expecting the LLM agent to decide which context to extract, we propose categorizing the sensing results into high-level user context descriptions. This approach simplifies the task for the LLM, making it easier to understand and utilize the provided context effectively. In particular, we carefully analyze the sensing data from the LifeSnaps [52] dataset and categorize them into the following three categories: Activity Level, Stress Level, and Sleep Quality.

Activity Level includes features such as 'calories burned', 'steps taken', 'distance traveled', 'lightly/moderately/very active minutes', and 'sedentary minutes'. *Sleep Quality* are characterized by 'sleep duration' and 'sleep efficiency'. *Stress Level* are monitored via 'stress score', 'root mean square of successive differences (RMSSD)', 'non-REM heart rate (nremhr)', 'resting heart rate,' 'beats per minute (bpm)', and 'mindful session participation'.

The above three categories are selected because they represent the most critical aspects of an individual's daily health and wellness, allowing us to create a holistic yet manageable user profile for the LLM to process. However, it's important to note that this categorization is not universally applicable. Different datasets or sensing sources may contain various types of information, necessitating alternative categorization methods. For instance, a dataset with more detailed nutritional or cognitive data might require additional categories for dietary health or mental sharpness. Therefore, the categorization method should be tailored to the specific dataset in use, ensuring that the most pertinent aspects of users' context are effectively captured and utilized.

Obtaining the trend index of Activity Level, Sleep Quality, and Stress Level: To help the agent establish a baseline for users and capture trends and variability, which may indicate meaningful changes in the user's state, each feature is also analyzed by computing a 7-day moving average and a standard deviation. A feature's trend is labeled as "high" if its current value exceeds one standard deviation above the moving average, indicating a significant increase compared to the usual trend. Conversely, a feature is classified as "low" if its value falls below one standard deviation under the moving average.

Output: The processed data is output as a structured file. For each feature, the file includes the current value, the calculated average, the standard deviation, and the trend index (high/moderate/low). This file is updated daily and serves as a key input for the Query Rewriting Module. This whole process allows WellMax to build a comprehensive profile of the user's physical and mental state, enabling more personalized, context-aware, and empathetic responses.

2.2 Query Rewriting Module

The Query Rewriting Module (QRM) is designed to leverage the processed sensing data to refine user queries so that the LLM agent's

Prompt Template: determine_refinement_goal

System Prompt:
You are an AI assistant that determines the goal for query refinement based on user summaries and queries... Based on these factors, infer the user's needs. For example, if the user has high stress levels and low activity...
Possible goals include: {Predefined Goals}
Follow the Examples:
Example 1:
Summary: {Example User Summary}
Query: {Example User Query}
Refinement Goal: {Example Goal}
{More Examples...}

User Prompt:
Given the following user summary and query, determine the primary goal for query refinement.
Summary: {User Summary}
Query: {User Query}

Refinement Goal:

Prompt Template: rewrite_query

System Prompt:
You are an AI assistant that rewrites user queries based on their summary and specific refinement goals. Follow **chain-of-thought** reasoning steps:
1. **Analyze User Data and Identify Key Factors**: Examine the user's profile data, including physical activity, ... Identify the key factors that influence the user's current needs (e.g., high stress levels, low activity, ...)...
2. **Determine the Refinement Goals**: ...Based on these factors, infer the user's needs. For example, if the user has high stress levels and low activity, they might need stress management...
3. **Rewrite Query**: Rewrite the user's query to address these inferred needs. Ensure that the rewritten query is concise, contextually relevant, and aligned with the user's emotional and physical state...
Follow the Examples:
Example 1:
Summary: {Example User Summary}
User Query: {Example User Query}
Refinement Goal: {Example Goal}
Rewritten Query: {Example Rewritten Query}
{More Examples}...

User Prompt:
Given the following user summary, original query, and refinement goal, please rewrite the query to incorporate the user's current state and needs.
User Summary: {User Summary}
Original Query: {User Query}
Refinement Goal: {Refinement Goal}

Rewritten Query:

responses are both contextually aware and closely aligned with the user's current physical and mental state. The QRM operates through a process inspired by the Chain-of-Thought (CoT) [45]prompting technique. Briefly, by guiding the model through intermediate reasoning steps, akin to human problem-solving processes, this technique allows the LLM to break down complex tasks into manageable sub-tasks, leading to more accurate and contextually relevant responses. The inference process is applied as follows:

• **Analyze User Data and Identify Key Factors**: In the first step, WellMax prompts GPT-4o to thoroughly examine the user's profile data, including the key indicators such as physical activity levels, sleep quality metrics, and stress levels. By analyzing this data, WellMax is able to gather and synthesize critical context, forming a comprehensive understanding of the user's current state.

• **Determine the Refinement Goals**: Once the user's profile data has been analyzed, WellMax then determines the primary query refinement goal. This step aligns the query refinement with the user's specific context and needs. In WellMax, we pre-define a set of potential goals based on common user needs and psychological principles: relieving stress, promoting physical wellness, improving productivity, enhancing mental clarity, and providing emotional support. WellMax uses contextual information from the profile analysis to select the most appropriate goals, ensuring the refinement process is both targeted and effective. To help WellMax choose the right goal, we design a set of few-shot prompting examples as the blue box above shows to help LLMs understand these goals and prioritize the user's current needs based on their context.

• **Rewrite the Query**: Inspired by Conversational Query Rewrite [31, 53], with the refinement goal identified, WellMax then proceeds to rewrite the user's original query. This step transforms the initial query into one that is not only more contextually appropriate but also more likely to yield a response that directly addresses the user's needs. For example, a query initially focused on task completion might be reframed to also consider stress reduction if the user's profile indicates high stress levels.

By applying the prompting technique shown in the black box above in this structured manner, our system effectively mimics human-like reasoning processes. This approach not only enhances the accuracy and relevance of the generated responses but also ensures that the interaction is personalized and aligned with the user's immediate needs and well-being.

A Running Example: Figure 4 illustrates a running example where Alice engages her personal agent with the query: *"How should I prepare for my presentation tomorrow?"*. The agent first extracts relevant sensing data, including Alice's heart rate, physical activity levels, and recent sleep patterns, from her wearable device. The data reveals that Alice has had moderate physical activity today but is experiencing elevated stress and reduced sleep duration. Given this context, the Query Rewriting Module focuses on relieving stress as well as preparing the presentation productively. The agent then rewrites Alice's query to: *"Considering my moderate activity levels, elevated stress, and low sleep quality, how should I prepare for my presentation to ensure it goes smoothly and I can manage my stress effectively?"*. This refined query is sent to the GPT-4o API, which generates a response that not only addresses presentation preparation but also offers strategies to manage stress.

3 EVALUATION

The focus of this paper is on the deep-dive into agents' replies and participants' feedback, as well as lessons learned from the experiment. We design experiments to (i): evaluate the effectiveness of our WellMax prototype and (ii): assess how well this sensor-in-the-loop design enhances user experience. **We obtained IRB approval to conduct user studies**.

3.1 Experiment Setups

Experiment Procedure. We emulate mobile users with different physical and physiological status (§3.1.1). Next, we emulate different use scenarios and compare the WellMax's response with the GPT4o-based LLM agent's output in each use case (§3.1.2). Finally, we carefully organize these two agents' responses into a questionnaire, gathering direct feedback from real users regarding the quality of the agent's responses, particularly in terms of context-awareness,

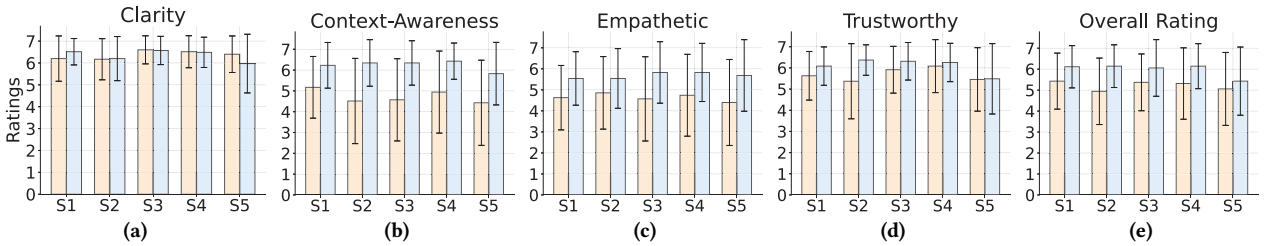

Figure 3: The clarity (3a), context-awareness (3b), empathetic (3c), trustworthy (3d), and ovreall ratings(3e) of each scenario (S1 ~ S5) with their standard deviation. The beige columns denote the baseline agent response overall score and the blue column is the response overall score of WellMax's response.

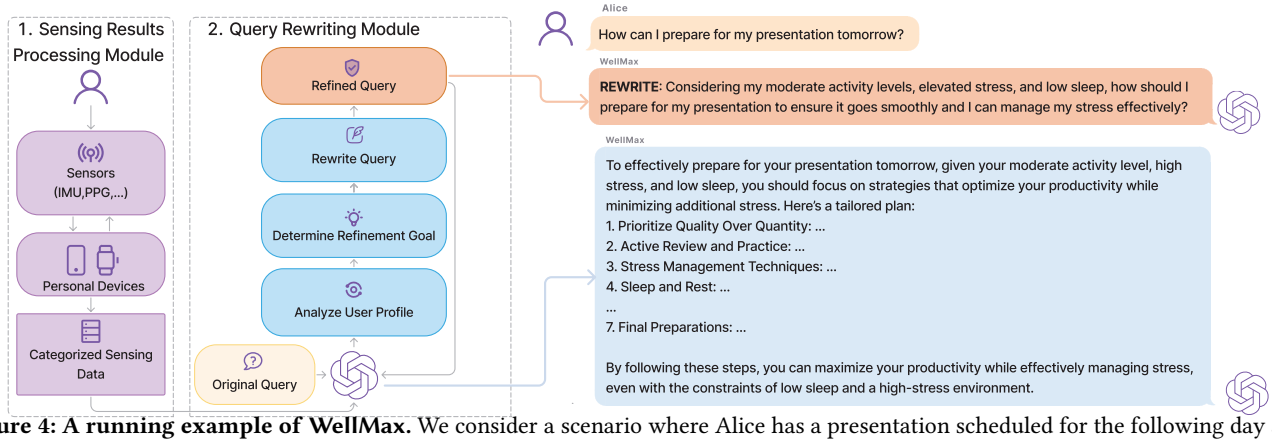

Figure 4: A running example of WellMax. We consider a scenario where Alice has a presentation scheduled for the following day and seeks guidance from the WellMax Agent.

empathy, relevance, as well as user preference. The users are also allowed to provide feedback on these LLM agent's responses.

3.1.1 User Emulation. We emulate users with different physical and physiological statuses using LifeSnaps dataset [52], which consists of real-world sensing data collected from wearable devices used by different individuals. LifeSnaps is well-suited for our study due to its diverse range of sensor data, including activity levels, sleep quality, and stress indicators, which closely align with the features our system processes. We select users from the LifeSnaps dataset based on varying combinations of activity level, sleep quality, and stress level. For example, we choose a user profile characterized by high activity level, high stress level, and low sleep quality in Scenario 4 to emulate a busy and exhausting day.

3.1.2 Use Scenario Selection. We carefully select five representative real-world use scenarios based on their relevance to common daily activities, as explained below.

1. Exercise Routine Planning – *I have availability for workouts in the mornings and evenings. What exercise routine should I start?* In this scenario, the agent helps create personalized exercise routines.

2. Healthy Eating Planning – *What should I have for dinner?* The agent offers personalized meal planning and healthy eating advice.

3. Presentation Preparation – *How can I prepare for my presentation tomorrow?* This use case involves assisting users in preparing for an upcoming presentation.

4. After Work Plan – *What should I do after work?* This use case focuses on helping users plan after-work activities.

Table 1: Summary of effect sizes (Cohen's d) and statistical significance across metrics and scenarios.

Metric	All	S1	S2	S3	S4	S5
Clarity	-0.03 (–)	0.34 (✓)	0.03 (–)	-0.06 (–)	-0.04 (–)	-0.33 (–)
Context	**0.73** (✓)	**0.72** (✓)	**0.73** (✓)	**0.93** (✓)	**0.83** (✓)	**0.56** (✓)
Empathy	**0.53** (✓)	**0.54** (✓)	**0.39** (✓)	**0.70** (✓)	**0.53** (✓)	**0.53** (✓)
Trust	0.30 (✓)	**0.45** (✓)	**0.56** (✓)	**0.36** (–)	0.13 (–)	0.02 (–)
Overall	**0.44** (✓)	**0.48** (✓)	**0.78** (✓)	**0.42** (✓)	**0.49** (✓)	0.17 (–)

Note: Each cell shows the Cohen's d effect size with its statistical significance in parentheses: ✓ for $p < 0.05$ and – for $p \geq 0.05$. Effect sizes (d) in bold are ≥ 0.36 (medium or large). Following Lovakov and Agadullina [33]: $d < 0.15$ = very small, 0.15–0.36 = small, 0.36–0.65 = medium, $d > 0.65$ = large. A p-value less than 0.05 indicates that a result is statistically significant. And a medium effect size can be interpreted as "obvious" to the observer.

5. Book Suggestion – *Can you suggest some books to read?* In this scenario, the agent recommends books that align with the user's current activity level and stress level.

We choose five different user queries, ranging from exercise planning to after-work leisure, so that each scenario would stress-test a different aspect of our sensor-in-the-loop approach. For instance, scenarios like "Exercise Routine" and "Healthy Eating" require fine-grained personalization based on physiological states (e.g., activity level), while "Presentation Preparation" and "After-Work Planning" focus on balancing mental fatigue and stress management. Finally, the "Book Recommendation" scenario serves as a lower-stakes, more open-ended task. By spanning these diverse use cases, we gain broader insights into how well our system adapts across various contexts and user needs.

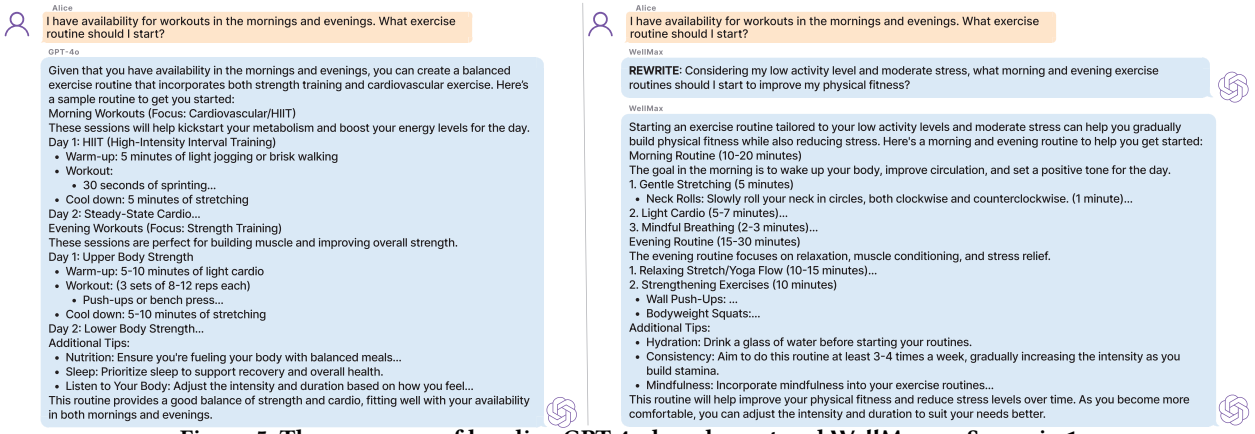

Figure 5: The responses of baseline GPT-4o-based agent and WellMax on Scenario 1.

In each use case, we feed the emulated mobile user's query, along with the sensor hints, to WellMax and record its response. For comparison, we also provide the same query to the GPT4o-based LLM agent without sensor hints. We then design a questionnaire by incorporating these two agents' replies and conduct a user study to gain direct feedback on the quality of the agents' responses.

3.1.3 User Study Design. We design a user study by recruiting 30 participants from UserTesting platform [42]. The questionnaire begins with a user physical and physiological state description obtained from the sensor data (§3.1.1), a user query (e.g., "*How should I prepare for my presentation?*") that requests agents to respond (§3.1.2), and a reply from either WellMax or the baseline agent.

The questionnaire shows the response from the baseline agent to participants. After the participant rating the response, the questionnaire will show the response from WellMax and ask these participants to report their preference. For fairness, we reordered the two responses in half of the questionnaires to exclude the order effect. Also, the questionnaire did not show participants the rewritten query. Hence the user is unaware that WellMax considers the emulated user's status when answering questions.

Following the 7-point scale, the participant is asked to rate the agent's response on (1): **Clarity** measures how easily users understand responses. (2): **Contextual Awareness** evaluates how well the responses align with users' current context. (3): **Empathy** assesses the degree of responses' empathy and emotional understanding. (4): **Trustworthy** indicates the users' confidence in the accuracy and reliability of the responses. (5): **Overall Satisfaction** reflects the users' overall satisfaction with the responses.

3.2 Experiment Results

3.2.1 Context-Awareness and Empathy. As illustrated by Figure 3b and 3c, WellMax consistently receives higher ratings than the baseline agent across all five use scenarios for context-awareness and empathy. Specifically, WellMax gets a context-awareness rating of 6.2 ± 1.2, which is significantly higher than the baseline agent's rating of 4.7 ± 1.9. To further validate our findings, we calculate the p-value through the Wilcoxon Signed-Rank Test and effective size through Cohen's d for paired samples. As Table 1 shows, the overall and per-scenario p-values for context awareness score are all less than 0.05 and the overall Cohen's d for context-awareness is 0.73, showing a significant difference between baseline and WellMax.

In terms of empathy, WellMax's responses are rated at 5.7 ± 1.5, surpassing the baseline's rating of 4.6 ± 1.9 by 1.1 points. The p-value and effect size analysis also show an obvious difference with $p<0.05$ and $d = 0.53$. Participants' feedback also indicated that WellMax excelled in understanding and responding to the user's current context, particularly in scenarios where physical and mental states played a crucial role in the appropriateness of the advice. Participants often praised WellMax for adjusting its responses based on real-time sensing data, making the advice feel more context-relevant and supportive.

For instance, in scenario 1, where the emulated user's activity level is low, the baseline agent suggests an intense exercise routine that includes activities like high-intensity interval training (HIIT), strength training, and cardio, as illustrated by the left part of Figure 5. While this routine might be suitable for a highly active individual, it is misaligned with the user's actual low activity level. One participant critically points out:

> *I think it did not align at all because it says the user's activity level is low, which means that they're not very used to being active. And then it pushes all these things on it like sprinting and stretching and cardio and yoga and bicep curls and all these things every single day... This would not be a good starting point for them.*

In contrast, the response from WellMax takes the user's low activity level into account. The response recommends a more gentle and realistic routine, starting with light stretching, light cardio, and mindful breathing in the morning, followed by relaxing yoga poses and body-weight exercises in the evening, as shown in the right part of Figure 5. The focus is on gradually building physical fitness and reducing stress. One participant noted:

> *The advice felt just right for the user's fitness level; it wasn't overwhelming, and it actually motivated a beginner to start."* Another added, "*I think this was a far more gentle approach, more realistic approach for someone in this situation...It would be great if you were really looking to improve your fitness, but bearing in mind you have low energy levels.*

Both experiment results and participants' feedback highlight the importance of context-awareness in creating effective and user-friendly interactions. WellMax's capability of adjusting its recommendations based on users' state makes the advice more actionable

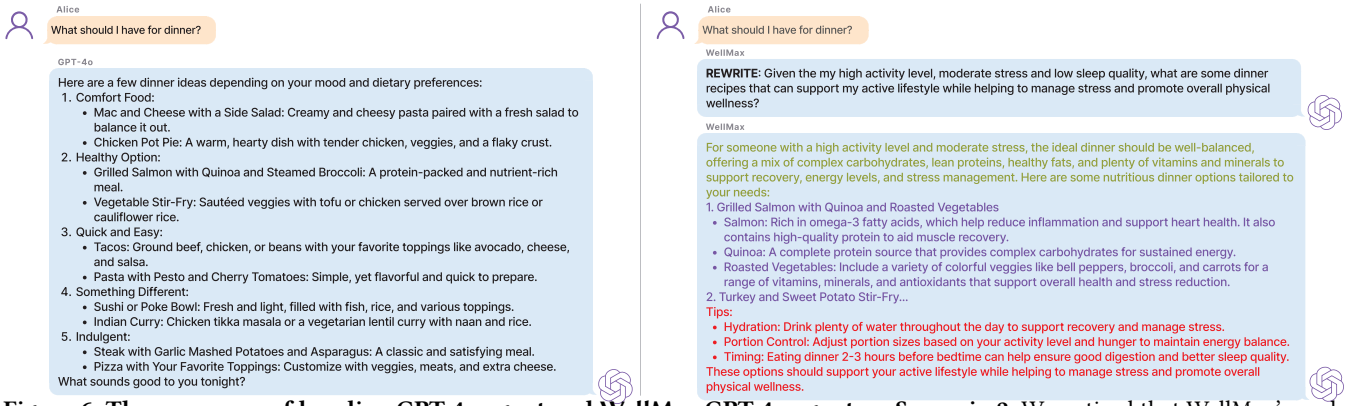

Figure 6: The responses of baseline GPT-4o agent and WellMax GPT-4o agent on Scenario 2. We noticed that WellMax's reply starts with an introductory paragraph that analyzes the user's needs based on his profile (highlighted in green) and ends with a conclusive paragraph offering additional tips (highlighted in red).

and also more empathetic, as it respects the user's limitations and provides a more compassionate approach to achieving their goals.

3.2.2 Trustworthiness. Figure 3d compares the average trustworthiness rating of responses from WellMax and baseline agent across five different scenarios. We find that WellMax is perceived as slightly more trustworthy compared to the baseline agent, with an average rating of 6.1 ± 1.1 versus 5.7 ± 1.4 for the baseline across five use scenarios. Cohen's d also shows a positive effect, with $d = 0.3$.

To understand why the ratings of two agents are on par with trustworthiness, we carefully check user feedback and find that most participants think both agents consistently provide accurate and generally helpful advice, which is foundational to trustworthiness. They also notice that the baseline agent, despite lacking personalization, can still derive responses that are logical.

The user feedback also reveals why WellMax's response is slightly preferred, particularly in how it personalizes advice to the user's specific context. Specifically, in Scenario 2, WellMax surpasses the baseline over 1 point in average with a $d = 0.56$ and p-value < 0.05. The emulated user has a high activity level and low sleep quality. The user query is: "*What should I have for dinner?*" The baseline LLM agent presents a variety of dinner options but fails to connect these choices with the user's physical condition, as illustrated by the left part of Figure 6. The response includes meal suggestions ranging from comfort foods like mac and cheese to healthy options like grilled salmon, without providing any rationale for why these meals were appropriate given the user's activity level and sleep quality. One participant criticizes this response, noting:

> *It should say in the response that if you have this activity, then you should be eating like this. But it's not very explained out of how you should be eating and why you should be eating the way you should be eating because of your activity level.*

This lack of explanation and context makes the baseline response feel less tailored and less trustworthy. In contrast, WellMax offers a well-balanced and contextually appropriate set of dinner options that are directly linked to the user's high activity level and low sleep quality. As illustrated by the right part of Figure 6, the response provides specific recommendations, such as grilled salmon with quinoa and roasted vegetables, explaining how each component of the meal would support recovery, energy levels, and stress management. Participants praise this approach, with one noting:

> *I think that they're not like, oh you have to eat grilled chicken and broccoli every single day. I think it gives some good options that still have flavors and won't make you crave unhealthy things.*

This level of personalization and consideration for the user's preferences and needs enhances the perceived trustworthiness of the WellMax LLM agent. The participants' feedback demonstrates the importance of integrating real-time contextual data into AI-driven responses to enhance trustworthiness. When users see that the advice is not only logical but also tailored to their specific circumstances, they are more likely to trust the system. While the baseline LLM agent produces credible advice, the WellMax agent's ability to provide contextually relevant and personalized recommendations makes it feel more reliable and trustworthy.

3.2.3 Clarity. Clarity in this context refers to how easily users can understand and interpret the LLM's responses. Figure 3a shows the ratings of these two agents on their responses' clarity. We observe that the rating of WellMax's responses is generally similar to, and occasionally slightly lower than the rating of baseline LLM agent's responses. The slight difference in performance here can be attributed to the inherently well-structured and user-friendly nature of modern large language models like GPT-4o, which are designed to produce coherent and straightforward language.

The rating gap between these two agents is most noticeable in Scenario 5 where the emulated user asks the LLM agent for recommending books. We notice that the incorporation of the physical and physiological status of the emulated user leads to the suggestion of a specific book with a detailed description. In contrast, the baseline LLM agent provides a broader list of options, as shown in Figure 7. Based on the participant's feedback, we found that while some participants appreciated the detailed recommendation, others found it too restrictive, preferring a wider selection of books to choose from. Although the detailed description in WellMax's response is clear, it may have reduced perceived clarity by narrowing user options too much. This participant's feedback highlights the importance of balancing detailed, contextually rich information with flexibility in user choices. We believe future iterations of the

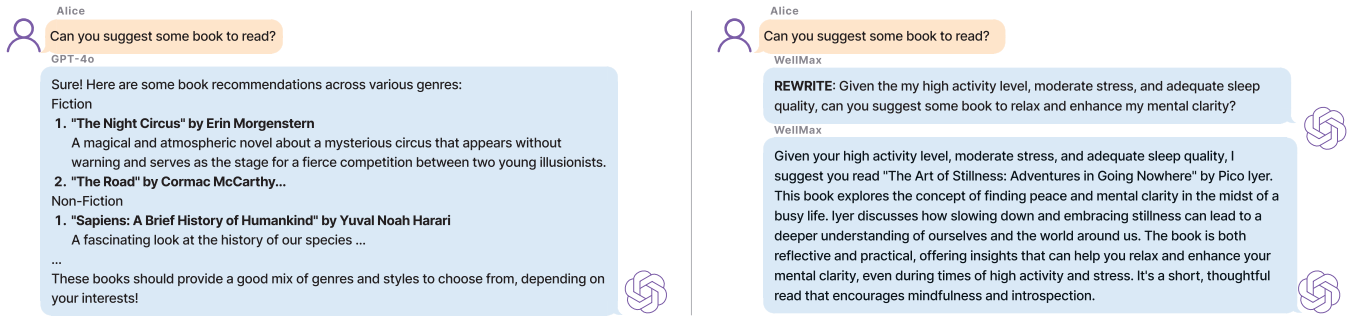

Figure 7: The responses generated by baseline GPT-4o agent and WellMax GPT-4o in Scenario 5.

WellMax agent should offer broader recommendations, each with detailed descriptions. This would maintain the richness of context provided by the sensing data while giving users more flexibility, enhancing both clarity and user satisfaction.

3.2.4 Overall Satisfaction. As shown in Figure 3e, the overall rating for WellMax (6.0 ± 1.3) is higher than the baseline agent (5.2 ± 1.6), further reinforcing the effectiveness of incorporating sensor data in generating responses that are not only empathetic but also more aligned with the user's context.

As illustrated in Table 1, the overall ratings of WellMax in Scenario 1 to 4 are statistically significant compared to the baseline, with p-value less than 0.05 and d across 0.42 to 0.78. While this effect is overall smaller than for Context-awareness and empathy, it is still in the medium range. The result demonstrates that WellMax's responses are overall perceived as more favorable and consistent compared to the baseline agent's responses.

3.2.5 Latency. We also analyze the latency associated with different stages of query processing, focusing on two metrics. **Time-to-First-Token (TTFT)** measures the time elapsed from when the system receives a query to when it generates the first token of the response. In our system, TTFT includes both goal determination and query refinement, collectively referred to as the *TTFT phase*. **Time Per Output Token (TPOT)** measures the average time taken to generate each output token during response generation.

To ensure accurate and robust latency measurements, we conduct five iterations of query processing using GPT-4o and report the average values. It should be noted that the reported latencies may vary due to service provider capacity and network conditions. The result shows that the average TTFT (Refinement Latency) is 1.290 seconds while the average TPOT (Time Per Output Token) is 0.017 seconds/token. These results indicate that query refinement introduces a measurable but controlled delay before response generation begins. Meanwhile, TPOT remains low, demonstrating GPT-4o's efficiency in token generation.

4 CASE STUDY ONE: WOULD SERIAL-POSITION EFFECT MATTER?

We noticed that when WellMax replies to a user query, its response begins with an introductory paragraph that analyzes the user's needs based on their profile. This is followed by a personalized suggestion that directly addresses the query, and finally, a conclusive paragraph offering additional tips.

For example, as illustrated in Figure 6 where the emulated user asks the LLM agent for dinner suggestions (Scenario 2). WellMax's

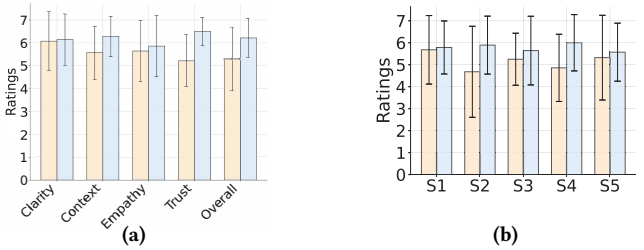

Figure 8: (a): Ratings over 5 metrics when swapping the introductory and conclusive paragraphs of Scenario 2. (b): The overall ratings bar-chart of case Study 2.

reply starts by recognizing the user's high activity levels and moderate stress, suggesting that a well-balanced dinner is crucial for supporting recovery, maintaining energy levels, and managing stress (highlighted in Green). It then provides a detailed list of nutritious dinner options tailored to the user's specific needs, such as grilled salmon with quinoa and roasted vegetables, and turkey and sweet potato stir-fry. Each option is carefully selected to include a mix of complex carbohydrates, lean proteins, and healthy fats, ensuring that the meal supports both physical wellness and stress management (highlighted in purple). The response concludes with additional tips on hydration, portion control, and meal timing, advising users to adjust portion sizes based on their activity level and to eat dinner 2-3 hours before bedtime to promote better sleep quality (highlighted in Red).

This raises a critical question: *how much of the user's satisfaction and retention is influenced by the beginning and the conclusion paragraphs versus the personalized content itself?* We pose this question because psychological research indicates that individuals tend to remember and focus more on the information presented at the beginning and end of a sequence, a phenomenon known as the serial-position effect (*a.k.a.*, primacy and recency effect) [7, 9]. In the context of Large Language Models (LLMs), this phenomenon is also relevant [18] and may influence how users perceive and evaluate generated responses.

To investigate whether the preference of WellMax over the baseline agent is due to the extra introduction and ending paragraphs it provided or the personalized suggestion alone, we manipulate these two agents' responses by deleting the introduction and conclusion paragraphs from WellMax's response in Scenario 2 and further adding these two paragraphs to the baseline's response. This allows us to test whether the relevance and personalization of the content generated by WellMax outweighs the influence of the introduction

and conclusion paragraphs on user preference. We then invite 30 participants to rate these responses.

Figure 8(a) shows the result. We observe that WellMax's new response (*i.e.*, contains only the personalized suggestion) remains the preferred choice among participants, achieving an overall rating of 6.2 ± 0.9, which is 0.9 higher than that of the modified baseline agent's response (5.3 ± 1.4) with added introductory and concluding paragraphs. These experiment results demonstrate that the relevance and personalization of the content generated by WellMax have a more substantial impact on user preference than the narrative structure, even when the baseline response is enhanced with introductory and concluding paragraphs. Participants consistently highlight the quality and relevance of the personalized suggestions in WellMax's response. One participant noted:

> *The meals that we provided here, they look very healthy and they all come with an explanation. So all the meals here seem to be tailored to the user's need what the body needs based on information provided by the user.*" Another participant remarked: "*I definitely feel like it did a really good job of thoroughly finding meals and multiple options that met what I needed, depending on how I went that day.*

These participants' comments suggest that while the serial-position effect exists in the context of LLM-generated responses, the content's relevance and direct alignment with the user's specific conditions play a more critical role in driving user satisfaction. This study reinforces the importance of query rewriting in generating personalized and context-aware responses. Moreover, the experiment results also validate the effectiveness of the sensor-in-the-loop design paradigm in enhancing AI-generated responses, leading to more meaningful and satisfying user interactions.

5 CASE STUDY TWO: WOULD MODEL SIZE MATTER?

The use of cloud-based LLM services like GPT-4o for processing sensory data may raise privacy concerns. One potential mitigation is to deploy small language models on the user's mobile devices, ensuring that data processing remains private. However, since model intelligence scales with model size [3, 24, 44], an important question arises: *"will using a smaller model compromise the performance of sensor-in-the-loop agent?"* To answer this question, we select Google's Gemma-2-2b-it, a 2-billion-parameter model, to evaluate how well WellMax performs under these constraints. We then extend our analysis (§5.2) with a focused experiment involving multiple small language models of various sizes to see how well they incorporate contextual data.

5.1 Performance of Gemma-2 WellMax Across Five Scenarios

Figure 8b shows the average overall scores of Gemma-2 WellMax across scenarios 1 to 5. We find that WellMax's responses achieve better overall ratings than the baseline agent's responses in Scenarios 2 (S2) and 4 (S4), while showing minor advantages in Scenarios 1 (S1), 3 (S3), and 5 (S5). To investigate why WellMax is perceived differently in these scenarios, we carefully analyze two agents' responses and the participants' feedback.

Observation One: The baseline Gemma-2 agent asks for more information before giving an answer and participants have mixed reactions to it across different scenarios. In Scenarios 1 (S1), 2 (S2), and 5 (S5), we notice that when the baseline agent is switched to the small model (Gemma-2), it tends to ask for additional information about the user's query rather than providing a direct response. In contrast, WellMax, which also uses Gemma-2, continues to deliver direct answers. For example, in S2 where the user has experienced a day of high physical activity and moderate stress, the baseline agent asks for additional details about cuisine preferences and available ingredients before offering detailed dinner suggestions. However, WellMax directly suggests meals tailored to the user's physical state, aiming to aid in recovery and stress management.

Although the baseline agent asks for additional information in all these three scenarios, we found participants' preferences for WellMax in S2 are significantly higher than in S1 and S5. While the underlying cause is still unclear to us, we suspect that the variation may be related to differences in the context of the query. Specifically, the emulated user asks *"What's for dinner"?* after a physically demanding day in scenario S2. Based on our personal experience, we believe that in such situations, participants tend to highly value quick, actionable advice without requiring additional interaction, as highlighted by the participant's feedback:

> *The meal options (from WellMax) were dictated and I believe that the user will like it. The other (baseline) response may get user confused between what to take. The user may not want to think too much about it.*

In S1, however, where the user with a low activity level and moderate stress asked for advice on starting a workout routine, the baseline's request for additional information such as fitness goals and current fitness level is more appropriate. Users recognize the importance of context for a personalized, effective workout plan. As one participant mentioned:

> *I appreciated that the system asked for more details about my fitness goals before suggesting a plan. I liked the list of things provided, but I feel like they're being a little more careful when asking follow-up questions to fit my needs.*

Observation Two: Gemma-2 WellMax fails to take into account both contextual personalization and user query in some user scenarios. In Scenario 3 (S3), the emulated user asks, *"How should I prepare for my presentation tomorrow?"* Like the baseline GPT-4o, the baseline Gemma-2 agent provides suggestions focused solely on the presentation itself, offering advice on content preparation, rehearsal techniques, and time management. The response from WellMax agent based on Gemma-2 focuses predominantly on stress relief and sleep improvement strategies, reflecting the user's high stress levels and poor sleep quality as indicated in their profile. The plan includes recommendations for relaxation techniques and a sleep routine, with minimal emphasis on presentation preparation.

Since the GPT-4o-based WellMax provides recommendations for both relaxation techniques and presentation preparation, and these responses are well appreciated by participants (see Figure 4), we believe that the lack of suggestions on presentation preparation makes the Gemma-2-based WellMax's response less appealing – the mixed feedback from participants further confirms our hypothesis. While some participants value the focus on overall well-being,

noting that reducing stress and improving sleep could indirectly enhance presentation performance, others feel that the response lacks the direct, actionable advice they were seeking about the presentation itself.

Observation Three: The baseline Gemma-2 agent and WellMax exhibit similar performance to their GPT-4o counterparts in other user scenarios. In Scenario 4 (S4), the emulated user asks, *"What should I do after work?"* after a day of very high physical activity, high stress levels, and poor sleep. Similar to the baseline GPT-4o agent, the baseline Gemma-2 agent's response offers broad suggestions such as socializing, reading, or cooking, which is generic and lacks the specificity needed to effectively address the user's immediate and actual concerns. In contrast, like GPT-4o WellMax, the Gemma-2 WellMax's response provides precise advice tailored to the user's high stress and fatigue. It suggests activities focused on relaxation and recovery, such as deep breathing exercises and gentle yoga, along with recommendations for an evening routine to support overall well-being.

5.2 Comparison with Other SLMs

We extended our investigation in Scenario 4 by testing three additional on-device small language models of various sizes: **(1) Phi-3.5-mini-instruct (3.8B)[1]**, **(2) Qwen-2.5-1.5B-instruct[49]**, **(3) stablelm-zephyr-3b[5]**, along with **Gemma-2-2b-it**. In this scenario, the emulated user has *very high physical activity, high stress, and poor sleep* and asks a simple question: **What should I do after work?**

Table 2 summarizes the rewritten queries and salient points from their answers. Below, we discuss key observations.

(1) **Goal-Informed Rewrites.** Despite the same original question, each LLM infers a slightly different *goal* from the sensor data. For instance, Phi and Qwen reframe the user's objective around "productivity," while stablelm emphasizes "stress management." Gemma takes a balanced approach, incorporating both relaxation strategies and physical activity.

(2) **Response Capability and Detail.** Although all models rewrite the query to reflect fatigue and stress, their final responses vary in specificity and empathy. Qwen briefly acknowledges poor sleep but offers mostly generic tips. In contrast, Gemma and stablelm use calmer, more empathetic language aligned with stress relief, yet stablelm lacks broader lifestyle suggestions.

(3) **Consistency with User Context.** The qualitative analysis notes that Gemma best aligned with the user's combination of fatigue and anxiety, providing multi-tiered relaxation. Phi delivers the longest list of ideas, though some are repetitive and aimed more at productivity than unwinding.

6 TAKEAWAYS

We summarize our takeaways from this study below.

(i): Personalization and Contextual Relevance. One standout feature of WellMax's responses is the ability to incorporate sensing data to personalize the interactions. Users remarked that WellMax's responses "take into current user context," making the interactions more contextual and personalized to their needs. For example, in the Presentation Preparation Scenario(S3), WellMax's responses not only suggest presentation techniques but also provide stress management strategies and sleep tips, which are crucial for creating meaningful and supportive user experiences.

(ii) Sensor Data as a Catalyst for Emergent Contextual Reasoning. A key observation from user studies is that WellMax 's enhanced context-awareness and empathy from considering seemingly straightforward sensor signals, such as daily step counts or stress levels into the prompt. In the Exercise Routine Planning Scenario(S1), even without explicit "if-then" rules linking "low activity" to "gentle exercise," the model spontaneously adjusts its recommendations to suit the user's current state. For instance, rather than planning heavy workouts for a low-activity user, WellMax proposes gradual, beginner-friendly routines that participants consistently describe as "more realistic" and "motivating." This emergent behavior underscores how minimal sensor cues can trigger the model to reevaluate user needs, leading to deeper reasoning about suitable actions, a more empathetic tone, and personalized advice. In turn, users feel better understood and supported. These results validate our claim that sensor data not only grounds the system's suggestions in real-time context but also unlocks richer, more adaptive responses beyond standard LLM outputs.

(iii): Detailed and Informative Responses. The integration of sensing data allows the LLM agent to provide more detailed and informative responses. Users appreciated the depth and specificity of the information provided. For instance, in the Healthy Eating Planning Scenario(S2), WellMax's responses not only list healthy meals but also explain the nutritional benefits and reasons behind each suggestion. This comprehensive approach helps users make better-informed decisions and improves their overall satisfaction.

(iv): The Preference of Immediate Usefulness vs. Additional Context. The higher preference for WellMax in scenario S2 (§5) demonstrates the importance of immediate usefulness in scenarios where the query is directly linked to the user's current physical state and demands quick, practical solutions. In contrast, in scenarios like S1 and S5 (§5), where queries are more exploratory or require a deeper understanding of user preferences, the interactive approach of the baseline agent is better received. These results reveal the need to balance immediate usefulness with the ability to gather additional context, depending on the user's state and the nature of the query, when designing an LLM agent.

(v): Performance on different Small Language Models. All four SLMs successfully *rewrote* the user query in a sensor-informed manner, but slightly differ in *how* they rewrite user queries and structure final responses. In S4, some tilt more toward productivity tips, while others emphasize stress management. Although all models benefited from sensor-informed rewriting, model size and training style continue to shape the detail of final recommendations.

(vi) Implications for Next-Generation AI Agents. Our findings suggest that sensor-driven LLM architectures like WellMax can serve as a blueprint for broader applications of context-aware and personalized AI. While we focused primarily on wearable data (e.g., activity, sleep, stress), this framework naturally extends to additional sensor inputs such as environmental metrics (air quality, temperature) or behavioral signals (social media usage, time-management data). By systematically fusing diverse inputs within an LLM's prompt-rewriting module, future AI agents could offer more holistic, domain-agnostic support—ranging from personalized health coaching to proactive lifestyle recommendations. In

Table 2: Evaluating the performance of other four small language models in Scenario 4.

Model	Rewritten Query and Core Reply Points
Phi-3.5-mini	**Rewritten:** *"Considering my very high activity levels, high stress and low sleep efficiency, what productivity strategies can I implement after work..."*
	Reply Highlights: Focuses on *productivity* with stress management tips, e.g., setting boundaries, trying structured relaxation (mindfulness, digital detox), and adopting better sleep hygiene. Offers a fairly comprehensive, list-based approach.
Qwen-2.5-1.5B	**Rewritten:** *"How can I enhance my productivity after work, considering high activity levels and low sleep quality?"*
	Reply Highlights: Centers on *task prioritization*, SMART goals, limiting distractions, taking breaks. Mentions "get enough sleep," but minimal direct reference to high stress; suggestions remain somewhat generic.
stablelm(3b)	**Rewritten:** *"What stress management techniques and relaxation exercises can I do after work..."*
	Reply Highlights: Concentrates on *relaxation routines*: aromatherapy, yoga, mindfulness, deep breathing. A few abrupt cutoffs in the reply; mostly stress-focused with limited discussion of daily scheduling or productivity.
Gemma-2-2b-it	**Rewritten:** *"Given my high activity levels, high stress and low sleep quality, what relaxing activities can I do after work to help manage my stress and improve my well-being?"*
	Reply Highlights: Provides *tiered relaxation ideas* (quick vs. moderate vs. longer routines), explicit tips on winding down, mindful hobbies, and customizing an evening routine. Emphasizes "experiment with what you enjoy." Balances stress relief and physical activity.

this sense, our work provides a foundation for evolving LLMs into robust multimodal systems that actively adapt to user contexts and operate seamlessly in real-world settings.

Future Research Directions. While conducting this research, we also identified several critical yet unexplored challenges with sensor-in-the-loop LLM agents.

(i): Coarse-grained vs. fine-grained sensor information. In this work, we use coarse-grained, averaged data summaries from LifeSnaps dataset as LLM's input. However, commercial wearable devices like the Apple Watch are capable of recording sensor data at tens of milliseconds of granularity, such as heart rate. Fine-grained data can offer a more detailed and accurate user profile representation. For instance, while an LLM might infer that a user with 95% sleep efficiency had a good night's sleep, fine-grained data could reveal that the user woke up for five minutes every two hours, indicating poor sleep quality. However, LLMs may struggle to accurately extract and interpret relevant insights from fine-grained sensing data. In what situations do we require fine-grained sensor data, and how can we efficiently process this information worth further exploration.

(ii): What if sensor results are inaccurate? We assume that the mobile user's physical and physiological activities inferred from sensor data are accurate and focus on how WellMax can utilize this information to enhance the response. However, in reality, the sensor readings can be erroneous due to interference, algorithmic errors, and behavior artifacts [30, 35, 57]. For example, GPS sensors can be inaccurate indoors or in densely populated urban areas due to signal obstructions. Likewise, the algorithms used to interpret sensor data may not be able to handle variations across individuals, activities, or environmental conditions. These cases would result in inaccurate physical activity descriptions, which may negatively impact LLM's response.

(iii): Should we feed LLM the time series sensor data or textual sensor result? A natural follow-up question is whether we should provide the LLM with raw sensor readings or continue using the textual sensor data as we currently do. The raw sensor readings are contextually rich, allowing LLM to learn detailed patterns and potentially identify errors. However, understanding raw sensor data from wearables poses a grand challenge to LLM due to its high dimensionality, which usually requires the LLM to understand individual data points and their dynamic patterns over time. Current research in both CPS [48] and healthcare [26] is already exploring the application of LLMs to time series.

(iv): Addressing Privacy and Ethical Implications of Sensor Data in LLMs. While GPT-4o offers advanced reasoning capacity and more nuanced responses, sending personal sensor data to the cloud would raise significant privacy concerns, as users may be concerned about sharing sensitive health information over the cloud. In contrast, Gemma-2, which can be deployed locally on mobile devices, provides a promising privacy-focused alternative, minimizing data exposure and offering users more control over their personal information. However, as our case study II (§5) shows, the small language model still suffers on specific tasks. So how to improve its performance is worth further exploration.

(v): Generalization and Scalability. Designing sensor-in-the-loop LLM agents inherently requires handling evolving user contexts. Currently, WellMax classifies user states into three main categories. While this categorization is straightforward for controlled integrations, it may not capture every real-world nuance. However, WellMax's *modular architecture* allows for incremental additions without a full modification. Once a new category is defined, the pipeline can incorporate it to refine queries accordingly. Furthermore, WellMax leverages *few-shot prompting* to flexibly handle novel contexts by learning from previously encountered scenarios. As a next step, we envision data-driven methods that can automatically discover and categorize latent user states. This direction aims to reduce manual overhead and enable WellMax to generalize seamlessly across diverse real-world settings.

(vi): Subjective vs. Objective Metrics. In typical text generation tasks, metrics such as ROUGE or BLEU evaluate similarity to a reference text. However, our personalized scenarios lack a single "correct" response. Consequently, we rely primarily on human evaluations—assessing clarity, context-awareness, empathy, trustworthiness, and overall satisfaction—to capture the nuances of sensor-informed interactions. It is worth noting that recent research has explored the use of LLMs as judges for open-ended tasks [17, 56], providing partial objective evaluations that correlate well with human judgments. Such approaches offer promising chances for complementing our subjective assessments, although they are not yet fully mature.

(vii): Prolonged User Study. We currently employ publicly available data to simulate user contexts due to the practical challenges of recruiting participants with specific physiological profiles and collecting consistent in-the-wild sensor data. However, we acknowledge that this approach cannot fully capture the dynamic and complex nature of real-world user states. As a future direction, we plan to conduct studies with recruited users to gather authentic sensor data over extended periods. Such studies will enable us to evaluate how sensor-informed responses evolve in real-world settings and further refine WellMax to better adapt to genuine user experiences.

7 RELATED WORKS

7.1 Understanding Sensor Data

Recent works explored the applications of LLMs to sensor data, focusing tasks such as trajectory recognition [51], and human activity recognition [19, 23, 32]. These demonstrated the potential of LLMs to generalize across different sensor data, achieving high levels of accuracy and understanding. However, many of these researches centered on sensor data recognition and classification, often treating the data in isolation from the broader context.

Penetrative AI [48] explores expanding LLMs' ability to interpret sensor data directly for various cyber-physical systems (CPS) applications, like activity recognition and heart rate detection. They examine how LLMs can leverage embedded world knowledge to process textualized and raw sensor data, enabling general perception and reasoning tasks within CPS contexts. In contrast, WellMax uniquely integrates real-time physical and physiological data to dynamically rewrite queries based on user-specific states, enhancing response quality through personalization. This approach ensures responses are contextually aware and empathetic, advancing user-centered conversational agents.

Some other researchers propose leveraging sensor data not only for recognition but also to enhance the user experience through context-aware applications. For example, Fang et al. [14] focuses on improving users' understanding of personal health information by integrating physiological data from wearable devices. In addition, Chen et al. [12] enables users to interact with the LLM-based system to create personalized context rules, which enhances the adaptability of context-aware IoT systems. Also, some works focus on leveraging LLM to perform high-level reasoning by integrating and interpreting complex sensor data for human activity understanding [21], health monitoring [26, 38, 50], enhancing mental health support [22, 37, 47] and analyzing socioeconomic impacts on environmental metrics [20]. Similarly, our work goes beyond recognition and classification by actively integrating sensing data to provide personalized and context-aware AI agent's responses.

7.2 Prompt Engineering

While sensor data provide enriched context, effectively leveraging this data needs techniques like prompt engineering which ensures accurate and relevant interactions between users and LLM agents. As an effective technique in PE, few-shot prompting allows LLM to generalize from limited examples provided within the prompt itself. Brown [11] demonstrated such capacity on GPT-3 and Gao et al. [15] further showed that even smaller models can benefit from this technique. Yu et al. [53] applied few-shot prompting to reformulate concise queries to contextual-aware queries.

Among various few-shot learning techniques, Chain-of-Thought (CoT) prompting[27, 45] gets attention as a method for eliciting complex, multi-step reasoning in LLMs. CoT prompting asks the model to generate intermediate steps when solving problems, thus improving its reasoning abilities. By imitating the human inference process, it enables LLMs to break down complex tasks into solvable substeps, thus reducing the errors and hallucinations of the outputs.

Our work explores some of these techniques to enhance the LLM's understanding of sensor-related contexts and we believe there is considerable potential for incorporating additional methods in future iterations.

7.3 LLM-based Personal Agent

Recent advancements in LLMs significantly improved the capabilities of personal assistants [29]. These models have demonstrated remarkable abilities in question-answering, task completion, and natural language understanding. However, as noted by [10, 25, 54], current LLM-based assistants often struggle with understanding user context and intent, particularly when queries are ambiguous.

To address these issues, Ma et al. [34] introduced a prompt-based task decomposition method that breaks down complex user queries into simpler sub-tasks, allowing users to explore these based on their preferences. However, this approach requires users to input personal context through a graphical user interface. Additionally, Zhang et al. [55] developed a demo that combines an on-device LLM with sensing data like screen text and user-reported questionnaires to enhance personalization.

Unlike [34], which requires manual input to specify a personal context, and [55] relies on user-reported questionnaire data, WellMax directly integrates enriched sensing data from wearable and mobile devices into the decision-making process of LLMs. It allows WellMax to implicitly and automatically enhance the response based on the user's physical and physiological context, providing more context-aware and empathetic interactions without requiring tedious user input. In addition, we provided extensive and comprehensive user studies and case studies across different experimental settings, highlighting the potential benefits and pitfalls of this sensor-in-the-loop approach in future LLM agents.

8 CONCLUSION

In this work, we demonstrated how integrating sensor data into LLM agents can yield more personalized and context-aware responses. By tracking user states via sensing data, WellMax can provide advice better aligned with immediate needs and long-term goals. At the same time, a balanced approach is essential, ensuring that recommendations remain both empathetic and practical. Looking ahead, we believe the success of future LLM agents will depend on their ability to interpret diverse sensor inputs and adapt dynamically to shifting priorities—whether users seek holistic well-being tips or task-specific solutions. As sensor-rich environments become more common, the next generation of conversational AI will likely be defined by how effectively it weaves together real-time context, user preferences, and adaptive reasoning to deliver truly meaningful and supportive interactions.

ACKNOWLEDGMENTS

We sincerely appreciate the insightful feedback from the anonymous reviewers and the shepherd. This material is based upon work supported by the National Science Foundation (NSF) under Grant No.2337537 and No.2441601. We also thank the University of Pittsburgh Center for Research Computing (CRC) for providing us the computing resource. Specifically, this work used the HTC and GPU cluster, which are supported by National Institutes of Health (NIH) under Grant S10OD028483 and NSF award number OAC-2117681.

REFERENCES

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*. PMLR, 265–279.

[4] Stability AI. 2024. Stable Assistant. https://stability.ai/stable-assistant. Accessed: 2024-09-10.

[5] Stability AI. 2024. stablelm-zephyr-3b. https://huggingface.co/stabilityai/stablelm-zephyr-3b.

[6] Apple Inc. 2024. Apple Watch. https://www.apple.com/watch/. Accessed: 2024-08-29.

[7] Solomon E Asch. 1946. Forming impressions of personality. *The journal of abnormal and social psychology* 41, 3 (1946), 258.

[8] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. 2015. Physical human activity recognition using wearable sensors. *Sensors* 15, 12 (2015), 31314–31338.

[9] Alan D Baddeley and Graham Hitch. 1993. The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition* 21 (1993), 146–155.

[10] Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with chatgpt. *arXiv preprint arXiv:2402.02136* (2024).

[11] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[12] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. From Gap to Synergy: Enhancing Contextual Understanding through Human-Machine Collaboration in Personalized Systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[13] Elena Di Lascio, Shkurta Gashi, Juan Sebastian Hidalgo, Beatrice Nale, Maike E Debus, and Silvia Santini. 2020. A multi-sensor approach to automatically recognize breaks and work activities of knowledge workers in academia. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–20.

[14] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. Physiollm: Supporting personalized health insights with wearables and large language models. *arXiv preprint arXiv:2406.19283* (2024).

[15] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).

[16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[17] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024).

[18] Xiaobo Guo and Soroush Vosoughi. 2024. Serial position effects of large language models. *arXiv preprint arXiv:2406.15981* (2024).

[19] Aritra Hota, Soumyajit Chatterjee, and Sandip Chakraborty. 2024. Evaluating Large Language Models as Virtual Annotators for Time-series Physical Sensing Data. *arXiv preprint arXiv:2403.01133* (2024).

[20] Zhizhang Hu, Yue Zhang, Ryan Rossi, Tong Yu, Sungchul Kim, and Shijia Pan. 2024. Are Large Language Models Capable of Causal Reasoning for Sensing Data Analysis?. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*. 24–29.

[21] Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2024. LLaSA: Large Multimodal Agent for Human Activity Analysis Through Wearable Sensors. *arXiv preprint arXiv:2406.14498* (2024).

[22] Sijie Ji, Xinzhe Zheng, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. 2024. MindGuard: Towards Accessible and Sitgma-free Mental Health First Aid via Edge LLM. *arXiv preprint arXiv:2409.10064* (2024).

[23] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? *arXiv preprint arXiv:2403.02727* (2024).

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[25] Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Aligning Language Models to Explicitly Handle Ambiguity. *arXiv preprint arXiv:2404.11972* (2024).

[26] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866* (2024).

[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[28] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702* (2023).

[29] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).

[30] Suwen Lin, Xian Wu, Gonzalo Martinez, and Nitesh V Chawla. 2020. Filling missing values on wearable-sensory time series data. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 46–54.

[31] Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational query rewriting with self-supervised learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7628–7632.

[32] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525* (2023).

[33] Andrey Lovakov and Elena R Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology* 51, 3 (2021), 485–504.

[34] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[35] Balz Maag, Zimu Zhou, Olga Saukh, and Lothar Thiele. 2017. SCAN: Multi-hop calibration for mobile sensor arrays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–21.

[36] Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Seyed Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. 2024. Openelm: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

[37] Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *arXiv preprint arXiv:2403.10779* (2024).

[38] Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. *arXiv preprint arXiv:2403.19857* (2024).

[39] Haroon Rashid, Sanjana Mendu, Katharine E Daniel, Miranda L Beltzer, Bethany A Teachman, Mehdi Boukhechba, and Laura E Barnes. 2020. Predicting subjective measures of social anxiety from sparsely collected mobile sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[40] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[42] UserTesting. [n. d.]. UserTesting Human Insight Platform. https://www.usertesting.com.

[43] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, Markus Gross, and Christian Holz. 2022. Affective state prediction from smartphone touch and sensor data in the wild. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[44] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[46] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[47] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[48] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.

[49] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[50] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. DrHouse: An LLM-empowered Diagnostic Reasoning System through Harnessing Outcomes from Sensor Data and Expert Knowledge. *arXiv preprint arXiv:2405.12541* (2024).

[51] Huanqi Yang, Sijie Ji, Rucheng Wu, and Weitao Xu. 2024. Are You Being Tracked? Discover the Power of Zero-Shot Trajectory Tracing with LLMs! *arXiv preprint arXiv:2403.06201* (2024).

[52] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data* 9, 1 (2022), 663.

[53] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1933–1936.

[54] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[55] Shiquan Zhang, Ying Ma, Le Fang, Hong Jia, Simon D'Alfonso, and Vassilis Kostakos. 2024. Enabling On-Device LLMs Personalization with Smartphone Sensing. *arXiv preprint arXiv:2407.04418* (2024).

[56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

[57] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.