GLOSS: Group of LLMs for Open-ended Sensemaking of Passive Sensing Data for Health and Wellbeing

AKSHAT CHOUBE, Northeastern University, USA
HA LE, Northeastern University, USA
JIACHEN LI, Northeastern University, USA
KAIXIN JI, RMIT University, Australia
VEDANT DAS SWAIN, Northeastern University, USA
VARUN MISHRA, Northeastern University, USA

The ubiquitous presence of smartphones and wearables has enabled researchers to build prediction and detection models for various health and behavior outcomes using passive sensing data from these devices. Achieving a high-level, holistic understanding of an individual's behavior and context, however, remains a significant challenge. Due to the nature of the passive sensing data, *sensemaking* — the process of interpreting and extracting insights – requires both domain knowledge and technical expertise, creating barriers for different stakeholders. Existing systems designed to support sensemaking are not open-ended or cannot perform complex data triangulation. In this paper, we present a novel sensemaking system, *Group of LLMs for Open-ended Sensemaking (GLOSS)*, for open-ended sensemaking capable of performing complex multimodal triangulation to derive insights. We demonstrate that GLOSS significantly outperforms commonly used Retrieval-Augmented Generation (RAG) technique, achieving *87.93%* accuracy and *66.19%* consistency compared to RAG's *29.31%* accuracy and *52.85%* consistency. Furthermore, we showcase the promise of GLOSS using four use cases inspired by prior and ongoing work in UbiComp and HCI communities. Finally, we discuss the potential of GLOSS, the broader implications, and the limitations of our work.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Ubiquitous computing**.

Additional Key Words and Phrases: sensemaking, large language models, mobile sensing, digital health & wellbeing, wearables

ACM Reference Format:
Akshat Choube, Ha Le, Jiachen Li, Kaixin Ji, Vedant Das Swain, and Varun Mishra. 2025. GLOSS: Group of LLMs for Open-ended Sensemaking of Passive Sensing Data for Health and Wellbeing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 76 (September 2025), 32 pages. https://doi.org/10.1145/3749474

1 Introduction

Humans continuously process vast amounts of information, constructing explanations to better understand situations or tasks at hand. This process is called *"sensemaking"*, a term widely used by researchers across various disciplines, including Human-Computer Interaction [83, 90], Organizational Studies [37], Information Science [29, 79], and others [8, 30, 51, 52]. It consists primarily of two iterative processes: (1) seeking, extracting, and filtering information, often referred to as the *information-seeking loop* (or the *foraging loop*), and (2) iteratively

Authors' Contact Information: Akshat Choube, choube.a@northeastern.edu, Northeastern University, Boston, USA; Ha Le, le.ha1@northeastern.edu>, Northeastern University, Boston, USA; Jiachen Li, li.jiachen4@northeastern.edu, Northeastern University, Boston, USA; Kaixin Ji, li.jiachen4@northeastern.edu, RMIT University, Melbourne, Australia; Vedant Das Swain, v.dasswain@northeastern.edu, Northeastern University, Boston, USA; Varun Mishra, v.mishra@northeastern.edu, Northeastern University, Boston, USA.

ACM 2474-9567/2025/9-ART76
https://doi.org/10.1145/3749474

developing and refining an understanding of the situation, known as the *sensemaking loop* [112]. Sensemaking is triggered by a *need for information* of a sensemaker (the individual or agent engaged in sensemaking), often expressed through a query. The information-seeking and sensemaking process continues until the sensemaker is satisfied with the developed understanding or explanations. Sensemaking is an innate process individuals engage in regularly – whether in making sense of conversations, interpreting trends in data, or solving problems. However, when the complexity of the data or the context exceeds an individual's expertise, external help—such as collaboration with others, tools, and frameworks—can enhance their ability to process and understand the information effectively. Data related to health, behavior, and well-being is often complex to interpret due to variability, multidimensionality, and sensitivity to context [5, 21]. Therefore, in recent years, researchers have made significant efforts to develop support systems [17, 86, 88] and tools [35, 62, 71] to assist in the sensemaking of health and behavioral data.

Passive sensing data collected using mobile phones and wearables offer a valuable source of behavioral and health information [1, 76]. Passive mobile sensing has demonstrated significant potential in monitoring and assessing various health and well-being outcomes, including depression [28, 101, 106], stress [25, 72, 74, 75, 92, 101], and anxiety [98]. Sensemaking of passive mobile sensing data can offer valuable insights to a broad range of stakeholders, including HCI researchers, behavioral scientists, clinicians, and individuals who actively engage in personal health tracking and informatics. Through sensemaking, HCI researchers and behavioral scientists can understand patterns in human behavior and can design effective interventions [62, 103]. Similarly, clinicians can make informed decisions and improve patient care through data-driven insights [2]. Likewise, sensemaking can assist self-health trackers in gaining a deeper understanding of personal well-being [24, 87], supporting disease management [70], and improving reflection [77, 96].

Researchers have been using passive sensing to build predictive models for specific behavioral outcomes (such as depression, stress, and physical activity), but a holistic, high-level understanding of an individual's well-being remains limited and challenging [2]. A key hurdle is the high dimensionality and data missingness of passive sensing data, particularly in its raw form, which makes it difficult to interpret [19, 91]. Furthermore, working with this type of data demands a specialized skill set, including expertise in handling and analyzing multimodal sensor data as well as proficiency in programming languages [48]. Drawing meaningful insights from passive sensing data often involves triangulating multiple sensor streams. For example, researchers might triangulate several passively sensed data streams, such as location, physical activities, and phone usage, to understand whether a participant had a productive day at work. People with limited coding skills and domain knowledge find it extremely difficult to interpret sensor data, let alone derive meaningful insights from it [18, 87]. Behavioral scientists and clinicians, who often have limited expertise in programming and analyzing sensor data, rely on domain experts to create dashboards and visualizations tailored to their needs. These dashboards, however, often capture only a limited subset of the dimensions of interest. Consequently, to obtain additional data or analysis, these stakeholders may need to seek further assistance from domain experts, making the process time-consuming and effort-intensive [48, 107]. Self-health trackers face similar challenges, as they are constrained by the predefined visualizations and insights provided by tracking systems/applications, often expressing a desire for deeper or more personalized information [6, 24, 54]. These barriers faced by various stakeholders in sensemaking and deriving insights from passive sensing data highlight the need for an open-ended, user-friendly, and accessible sensemaking system.

Current sensemaking and data exploration tools for passive sensing data are either not open-ended [23, 71] or often deal with simplified tasks like reporting statistical metrics of pre-calculated measures [35]. Stakeholders require a certain level of domain knowledge and expertise to effectively interact with and use these tools [16, 62]. For example, many existing systems work with aggregated or processed data [71], rather than directly using raw data from sensing streams, necessitating expertise in sensor data preprocessing. Additionally, current sensemaking systems often fall short in delivering customized information tailored to the varying needs of different stakeholders [14]. For instance, most sensemaking systems reduce individuals to quantified beings, failing to provide a holistic view of their health and wellbeing [96]. In recent years, Large Language Models (LLMs) have demonstrated impressive performance in "common sense" reasoning [27, 69], medical question answering [42, 95], and programming tasks [63, 65], making them a promising component in the design of sensemaking systems for health and behavioral data [35, 71, 96].

In this work, we introduce **GLOSS** (**G**roup of **L**LMs for **O**pen-ended **S**en**s**emaking), a novel sensemaking system designed for passive sensing data collected via phones and wearable devices. GLOSS uses a network of large language models that collaborate through rules and processes inspired by models of sensemaking in HCI [83] and Organizational Studies [51]. This fully automated system integrates function calling and code generation capabilities to retrieve and analyze passive sensing data. GLOSS is a query-based system where users[1] can seamlessly express their information needs by posing queries in natural language. GLOSS offers a scalable, automated, and off-the-shelf solution, eliminating the need for domain expertise or coding skills and making it highly accessible to diverse stakeholders. GLOSS can answer simple data retrieval queries like *"What was the last social media app this person used before 11 am?"* to more complex queries like *"Were there anomalies in how this person used their social media"*, along with reasoning questions like *"Why did this person self-report a high-stress event on Monday evening?"*. We present an evaluation of the GLOSS's underlying system on objective and subjective queries and compare it against a Retrieval Augmented Generation (RAG) technique. Following this, we demonstrate how GLOSS can drive a variety of outcomes and research methods: a chat-based system for researchers to query & analyze data, generating personal informatics narratives to support reflection, and even automated tasks like explaining anomalies and personalizing EMA prompts. We keep detailed evaluations for each of the use cases in future work.

Thus, we make the following **contributions**:

- We introduce GLOSS, a novel query-based sensemaking system that is open-ended, scalable, and extensible, designed to support diverse stakeholders in interpreting passive sensing data.
- We conducted technical evaluations benchmarking the performance of GLOSS, demonstrating promising results compared to the RAG technique.
- We present the application of GLOSS across four important use cases inspired by previous and ongoing work in the UbiComp and HCI communities.

GLOSS is the foundational step towards a fully open-ended and multimodal sensemaking system designed to work directly with raw sensor data, paving the way for numerous future applications. Our code for the GLOSS system is publicly available on GitHub[2] to enable other researchers to experiment with GLOSS on their own data and extend this work.

2 Related Work

Our work builds on prior research on passive sensing data from smartphones and wearables, the sensemaking process, and recent LLM-driven data exploration systems.

2.1 Barriers in Interpreting Passive Sensing Data

Mobile and wearable devices can collect large amounts of passive sensing data on a daily basis. Researchers leverage these longitudinal streams of passive sensing data to predict a variety of health and behavioral outcomes [7, 11, 44, 93]. Researchers can use these predictive models to drive context-aware interactive systems, just-in-time interventions, or health monitoring tools. Although there have been multiple breakthrough attempts

[1]Here, users could be to both humans or automated systems that need to interpret data through querying GLOSS. Throughout the rest of the paper, when we refer to "users," we include both.
[2]https://github.com/UbiWell/GLOSS

to incorporate mobile passive sensing data into human behavioral detection and personal health informatics (PHI) systems [67, 73], interpreting longitudinal passive data remains challenging for several reasons. First, processing and interpreting passive sensing data often requires domain knowledge and technical proficiency [107]. Many researchers from non-technical backgrounds, like psychology or behavioral science, might struggle when processing raw sensor data [36]. Second, the patterns of passive sensing data are often highly individualized [108] – different participants have different behavioral patterns, thus often requiring researchers to train personalized models for each participant [40, 105]. Different datasets also have varying formats and sensor streams available, often with significant amounts of missing data [19, 33, 45]. This inconsistency prevents researchers from developing generalized, off-the-shelf models that can be used without any additional training or fine-tuning [39]. Finally, most behavioral detection and PHI focus on a limited predefined health aspects (e.g., depression [66, 105], anxiety [98], physical activity counts [12]), which limits the ability of researchers to ask open-ended questions or triangulate information across different data sources . A system capable of responding to open-ended, natural language input allows researchers to explore more complex questions and gather insights that might not be immediately accessible in structured formats or require reasoning based on common sense knowledge [47, 68, 102]. These barriers are exacerbated for people interested in understanding personal health behavior through self-tracking. Attig et al. [6], found that data incomprehensibility was one of the important reasons of wearable activity tracking attrition. Likewise, Lazar et al. [54] argued that being able to understand data from smart devices increases motivation to keep using those devices. Thus, to assist in interpreting and understanding passive sensing data (sensemaking), researchers have built support systems and tools.

2.2 Sensemaking of Passive Sensing Data

Sensemaking as a concept gained popularity starting in the late 1970s [30, 79]. Researchers have expanded on this concept and applied it across various fields, such as organizational theory, education, decision-making, and human-computer interaction [29, 37, 51, 83]. The broad idea of sensemaking being an iterative process that involves information seeking, and making sense of information, remains consistent across these fields [112]. Sensemaking of passive sensing data is a complex and challenging task. In prior work, researchers have developed support systems and networks to share passive sensing data with others, facilitating discussions about their health. Puussaar et al. built a social network for self-trackers [85], while Coşkun suggested building social communities and groups for collaborative sensemaking [24]. While these approaches can assist in sensemaking, they often rely on the expertise of other people and are often not widely accessible.

Researchers have also explored innovative data visualization techniques to support users with the interpretation of passive sensing data. These approaches, however, do not take the additional step of performing the sensemaking process for the users. Choe et al. introduced a web-based visualization dashboard to support self-reflection and allowing users gather insights on their personal data, but they found that data interpretation was a key hurdle [18]. Adler et al., through an in-lab study with clinicians, showed that while visualization dashboards provide actionable insights for treatment, they often lack context and introduce personal biases. Mamykina et al. demonstrated that visualizations aimed at disease management for diabetes patients assisted in sensemaking but were considered complex to interpret due to the need to triangulate data across visualizations [70]. Likewise, Karahanoğlu et al. showed that self-health trackers get overwhelmed with the data presented to them if not properly curated [48].

Thus, visualizations-only approaches for passive sensing data are still considered hard to interpret by clinicians, behavioral scientists, self-health trackers, and even researchers. With the advent of Large Language Models (LLMs), researchers have started using LLMs' capabilities in common sense reasoning, question answering, and code generation to design systems that can do sensemaking on behalf of the users.

2.3 LLMs with Passive Sensing Data

Previous research has used LLMs to develop predictive models for various health and behavioral outcomes, including depression, stress, sleep quality, dementia diagnosis, and activity recognition [46, 50, 82]. As LLMs primarily process natural language data, these studies convert passive sensing data into textual representations. Researchers subsequently vectorize these textual representations, storing them in vector databases and employing Retrieval-Augmented Generation (RAG) techniques to retrieve relevant data to queries [58, 110]. Yang et al. introduced an AI doctor capable of retrieving sensor data and medical knowledge through RAG to suggest medical advice [109]. Furthermore, in some studies, researchers have fine-tuned LLMs using wearable data for tasks like sleep quality prediction [23] and activity-based question-answering [16]. Researchers have also built systems that can analyze passive sensing data and offer insights: Stromel et al. created LLM-based narratives of daily step count data to trigger self-reflection [96]. Likewise, Cosentino et al. [23] fine-tuned a large Medical LLM with sensor data to generate sleep and physical activity insights. Xiong et al. presented a tutorial on building context-aware LLM-based systems to command sensor actuators (e.g., a mobile z-arm in an apartment for assisted living) that users can operate via natural language commands [104]. In a complementary work to building such LLM-based systems, Li et al. [62] developed a prototype system called Vital Insight and investigated its applicability in assisting domain experts' sensemaking process. Some researchers have also argued that most existing PHI systems do not offer personalized insights [89], often reducing individuals to "quantified beings" [96].

Two closely related works are those by Fang et al. [35] and Merrill et al. [71]. Fang et al. presented an interactive system, PhysioLLM, that is capable of providing insights derived from physiological data. These insights, however, are limited to finding correlations between aggregated data (for instance, between hourly step counts and hourly heart rate) through precomputed correlation matrices. Merill et al. developed PHIA, an LLM-based system that answers user queries through code generation and Google search. PHIA follows a structured three-step (thought-act-observe) process to interpret data but is restricted to working with daily aggregated data stored in CSV files (e.g., daily total steps, daily wake-up time, etc.). The authors evaluated PHIA using template-based queries, thus restricting its adaptability to more open-ended inquiries. Additionally, PHIA's code generation is confined to basic statistical operations (e.g., mean, median, standard deviation, and correlations), making it less suited for answering complex queries that require deeper data integration and triangulation. We summarize the limitations of existing systems in Table 1.

Our proposed system, GLOSS, addresses these limitations: GLOSS retrieves, processes, and analyzes raw passive sensing data directly from databases rather than relying on pre-aggregated or pre-computed data and metrics. Additionally, GLOSS can generate complex code that goes beyond basic statistical operations, enabling integration across multiple data streams and facilitating data triangulation. Moreover, GLOSS does not require any model training or fine-tuning and can easily be extended to include various sensor streams. Thus, we believe GLOSS is the first fully open-ended sensemaking system that enables different stakeholders to analyze and interpret passive sensing data more effectively. We highlight these differences between prior works and GLOSS in Table 2.

3 System Design

GLOSS consists of a network of Large Language Models (LLM) designed to simulate the process of sensemaking in humans, drawing inspiration from the human sensemaking frameworks proposed by Klein et al. [51] and Pirolli & Card [83]. These frameworks incorporate two interconnected cyclic processes: (1) the *Information Seeking/Elaboration* loop focuses on seeking out new information and filtering, aggregating, or transforming it into an accessible and understandable format; and (2) the *Sensemaking/Reframing* loop involves analyzing and interpreting existing data, deciding whether additional data is required. If no further data is needed, it focuses on determining the most effective way to present the information

Table 1. Limitations of current systems, their descriptions, and example queries that those systems would fail to answer.

Limitation	*Description*	*Example Query*
Uses Aggregated Data	Some systems rely on data aggregated over hours or days, limiting their ability to provide fine-grained insights.	*What app was the user using around 8:30 pm yesterday? Was it a social media app*?
Limited Computation	Existing systems often generate code for simple statistical functions or use precomputed results, limiting their ability to perform advanced calculations.	*Did the user remain within one mile of their home throughout the day on 28th October 2024?*
Single Data Stream	Existing systems use a single or limited set of data streams, reducing the insights they can provide.	—
No Data Triangulation	Some systems use multiple sensor streams but lack the ability to triangulate data between them.	*What apps was the user using when their heart rate spiked over 120 bpm on the 17th of September?*
Requires Model Training	Some systems require model fine-tuning, limiting their adaptability and restricting their usability to specific domains.	—
Impersonal Insights	Many existing systems cannot tailor their responses to the specific needs of the users.	*Summarize the physical activity of the user on 25th August 2024, focusing on qualitative aspects.*
Requires Expertise	Some existing systems assist in sensemaking but still need passive sensing or programming expertise.	—

Table 2. A summary of the capabilities of related LLM-based sensemaking approaches, where ● indicates the presence of a particular capability in the given approach. ◐ in Multi data streams denotes the system uses multiple data streams but does not do complex data triangulations.

Model	**Open-ended**	**Raw passive sensing data**	**Multi data streams**	**Query-based interface**	**No model training**	**Code generation**	**Open source**
Narrating Fitness [96]	○	●	○	○	●	○	○
PH-LLM [23]	○	●	◐	●	○	○	○
Vital Insight [62]	○	●	◐	○	●	○	○
PhysioLLM [35]	○	○	◐	●	●	○	○
PHIA [71]	○	○	◐	●	○	●	○
GLOSS (ours)	●	●	●	●	●	●	●

We designed GLOSS to directly address the barriers in sensemaking that exist in prior works (Table 1). Thus, GLOSS works with high-dimensional *raw passive sensing data* from *multiple sensor streams* and is capable of generating *advanced code* to *process and analyze* the data. It can be used *off-the-shelf* (i.e., without any model training). GLOSS is a query-based system that can be used in human-facing tasks or automated background tasks. Moreover, GLOSS has *minimum learning barriers* for human-facing applications as users can express their information needs as natural language queries. Additionally, GLOSS can *personalize answers* based on an individual's needs.

3.1 System Components and Flow

We present the overall design of GLOSS in Figure 1. To illustrate how GLOSS operates, we first define the necessary terminologies in Table 3.

Table 3. Sensemaking Process Components.

User Query	The natural language query that expresses the user's information need and initiates the sensemaking process. A user can be a human stakeholder or another automated system querying GLOSS.
Action Plan	A high-level strategy to drive the sensemaking process with the aim to address the *user query*.
Information Request	A specific data retrieval request made to access databases or machine learning models available in GLOSS.
Memory	Memory contains all the data retrieved so far using *information requests* to develop an *understanding* of the user's query.
Understanding	The interpretation or the answer to the *user query* derived so far by the system based on data in the Memory. It also includes information on what additional data is needed to refine or expand this interpretation/answer.
Answer	The answer to the query retrieved from *understanding* and presented to the user according to their needs.

In GLOSS, multiple LLM agents interact with each other with a shared objective to interpret passive sensing data based on the *user query*. The sensemaking process begins when the user submits the *user query* expressing the information need (1).

3.1.1 Action Plan Generation. Once GLOSS receives a *user query*, the *Action Plan Generation Agent* formulates a high-level *action plan* to address the query using the system's access to available data and models. The *action plan* provides a high-level, step-by-step guide to follow in order to derive an answer for the *user query*. In the prompt to the agent, we detail information on available databases (sensor streams such as location data, Wi-Fi, Bluetooth, etc.) and models (e.g., a stress prediction model or an anomaly detection model). The shared information includes a description of the contents of the database and the device used to collect the data (phone or smartwatch). The agent then uses this knowledge to devise an appropriate course of action. If the query cannot be answered with the available datasets, the agent halts the sensemaking process and informs the user; otherwise, it continues the process. We show an example of an Action Plan when query requests data stream not currently supported by GLOSS in Figure 2 (a).

3.1.2 Information Seeking/Sensemaking Loop. Following the action plan generation step, the sensemaking process transitions into the Information Seeking/Sensemaking Loop, beginning with an empty *understanding* and *memory* (2). At the start of this loop, the Next Step Agent uses common sense reasoning capabilities of LLMs to evaluate whether the current *understanding* sufficiently addresses the *user query* based on *action plan* generated in the previous step. If it does, the process concludes; otherwise, it continues to the Information Seeking step.The Next Step agent also halts the process if the understanding contains information that the data couldn't be fetched

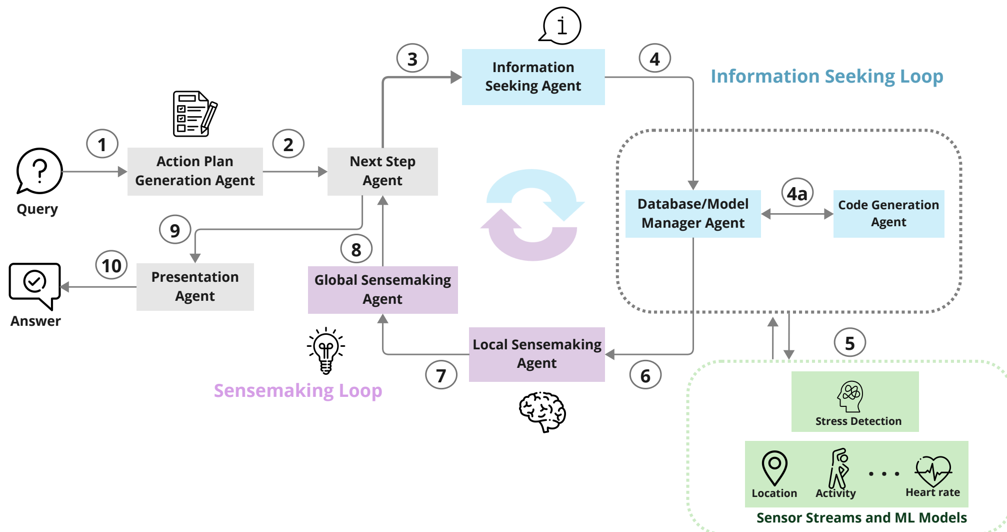

Fig. 1. GLOSS system design where different LLM agents interact according to prior sensemaking theories.

due to technical issues or data missingness. Since in the first iteration, *understanding* is empty and thus cannot answer the *user query*, the process moves to the information-seeking part of the loop (3).

In this loop, the Information Seeking Agent comes up with specific *information requests* with an aim to gather data relevant to developing a better *understanding* of the *user query*. These requests can involve fetching and triangulating data from multiple databases. The Information Seeking Agent makes these requests to Database/-Model Manager Agent (DM Agent) based on the *user query*, *action plan*, *memory*, and current *understanding* (4). The DM Agent retrieves the relevant prebuilt data retrieval and processing helper functions based on the information requests. For example, if an information request is "fetch me all GPS location values where activity is running," the DM agent retrieves all helper functions for location and activity databases. After retrieving helper functions, DM Agent shares these functions along with their parameters, and their return values with the Code Generation Agent (4a). The Code Generation Agent creates Python or bash code to process the data as per the request made by the DM Agent. It uses helper functions to retrieve the data from different databases relevant to the information request and generates additional code to process and triangulate data to fulfill the request. The agent makes multiple attempts to complete the request, attempting to resolve any programming errors on its own. The code generation capabilities of LLMs allow GLOSS to handle queries that involve basic to complex calculations, processing, and data triangulations. The Code Generation agent then executes the code that can directly access raw data from databases or ML models through helper functions (5). After processing the data and computing results, the Code Generation Agent sends these results to the DM Manager Agent, who then forwards them to the sensemaking phase of the loop (6).

In the sensemaking loop, the first step is local sensemaking. The results of the code generated by the Code Generation Agent to complete the information request can be Python data structures (like lists, dictionaries, etc.), and the Local Sensemaking Agent's role is to generate a natural language representation of these results. This step is crucial, as LLMs are more effective at processing data in natural language than in numerical formats or data structures [57, 97]. Following this, the *memory* is updated with the *information request* and the corresponding natural language representation of the answer. The process then advances to the global sensemaking step (7), where the Global Sensemaking Agent is responsible for refining the previous *understanding* of user query based on new information added to memory in the current iteration (in local sensemaking step). This agent reviews the *action plan*, *previous understanding*, and all the data fetched so far present in *memory* to come up with a refined *understanding* of the *user query*. Moreover, while updating the *understanding*, this agent is instructed to include information on any additional data that might be helpful in completing/strengthening the *understanding* in the subsequent iterations. Additionally, if in the current iteration, the data fetching and processing was unsuccessful due to technical issues or data missingness, the agent includes it in the *understanding*. In Figure 2 (b), we show an example *understanding* when we forcefully restricted GLOSS's access to raw data to create a data retrieval failure. After the global sensemaking step, the process goes back to the Next Step Agent again (8), completing one iteration of this loop. The loop continues until the Next Step Agent determines that the current *understanding* sufficiently addresses the *user query* or there is information on some technical failure or data missingness in the *understanding*.

3.1.3 Answer Presentation. Once the Next Step Agent decides to conclude the process (9), the *understanding* is handed off to the Presentation Agent. This agent is instructed to extract the *answer* to *user query* from the final *understanding* and present it to the user according to specific presentation instructions or formats. This agent tailors the presentation to meet the specific needs of different users, as they may require the information in various formats. For example, human users can provide instructions to highlight qualitative aspects of well-being, or automated systems can request data in JSON outputs.

Overall, GLOSS mimics the iterative information-seeking/sensemaking process described by prior theories on human sensemaking [51, 83]. It retrieves necessary data and processes it, refining its understanding of the query in each iteration and finally presenting the answer to the user based on their needs.

Action Plan
The query is regarding PPG (Photoplethysmogram) average value for `test006` on 2024-09-28. However, none of the provided databases contains PPG data. The available databases focus on activity, location, step count, heart rate, lock/unlock times, wifi connections, app usage, battery status, call logs, and stress levels. Since PPG data is not covered by any of these databases, the query cannot be answered with the given datasets.

(a)

Understanding
To determine the total duration of each activity done by `test004` on 2024-07-12, we accessed the activity database for records associated with this user. Unfortunately, no activity data was found for `user_id` 'test004' on the specified date of '2024-07-12'. As a result, it is not possible to calculate the total duration for each activity type, such as stationary, automotive, cycling, walking, and running, for that day. This absence of data could be due to the user not having any recorded activities on that date or an issue with data retrieval. If further data becomes available or if there is an alternative data retrieval method, it may be possible to revisit this query.

(b)

Fig. 2. (a) Action plan when query requests data not supported by GLOSS (b) Understanding when GLOSS has technical issues.

3.2 Implementation Details

We implemented GLOSS in Python using OpenAI, LangChain, and AutoGen libraries to create and manage interactions between various LLM agents. We used GPT-4o [81] as our Large Language Model and designed agent prompts following OpenAI's recommended best practices for prompt engineering [80]. In our GPT-4o API calls, we maintained the default hyperparameter values ($temperature = 1$ and $top_p = 1$). To ensure security and prevent potential harm to the system running GLOSS, we executed any code generated by the LLM models within a Docker container. Our GPT-4o deployment was HIPAA compliant, and OpenAI did not retain any queries for training their models. Moreover, we do not share large chunks of raw data with GPT as the code that GPT generates to access raw data is run in the local system running GLOSS, and then aggregated statistics are shared with the models. To prevent a situation where the information seeking/sensemaking process might run into an infinite loop, we set up a hard cutoff of five iterations, after which GLOSS presents an answer to the user derived from the latest *understanding*. We have included sample runs of GLOSS to different queries in the supplementary materials.

GLOSS supports multiple streams of passive sensing data while providing a scalable and modular framework for incorporating additional data streams. For our experiments and evaluations, we integrated several passive sensing data streams collected from iPhones and Garmin wearables. Details of the data streams are provided in Table 4. The raw passive sensing data was collected using an iOS application and stored in MongoDB databases. GLOSS can extract raw data directly from these MongoDB databases through some pre-coded helper functions. Our data collection protocol was approved by the Institutional Review Board (IRB) at our institution. We can easily extend GLOSS to include additional data streams or ML models by just providing information about the data stream and some helper functions to fetch and retrieve data. We show sample databases and helper function descriptions we provide to agents in Figure 3.

Table 4. Data sources, their descriptions, and frequency of collection.

Data streams	*Description*	*Sampling rate*
Location	Provides GPS coordinates (latitude, longitude, and altitude) from the phone.	1 minute
Activity	Provides activity categorizations such as stationary, walking, cycling, running, and automotive, using iOS's built-in activity detection.	Event-driven
App usage data	Provides open and close times of different iOS applications tracked through setting automations in the phone	Event-driven
Phone steps	Provides step counts, floors ascended/descended, and distance covered using phone sensors.	1 minute
Phone lock/unlock	Tracks phone lock and unlock events.	Event-driven
WiFi	Provides connected/not connected status and WiFi details (WiFi name and SSID) when connected.	1 minute
Call logs	Records incoming, outgoing, and missed calls.	Event-driven
Phone battery	Monitors changes in the phone's battery status.	Event-driven
Garmin steps	Tracks step counts from a Garmin smartwatch.	1 minute
Garmin heart rate	Provides heart rate in bpm (beats per minute) from a Garmin watch.	30 seconds
Predicted Stress Levels (ML model)	Prediction of physiological stress (between 0:low stress and 1:high stress) using IBI and heart rate signals from Garmin watch based on the open-source model by Mishra et al. [75].	30 seconds

App Usage Database:
Info: Contains app usage data, including app names, open and close times, and durations.
Device: Phone

Garmin Stress ML Model:
Info: Contains physiological stress predictions from ibi data recorded from the Garmin smartwatch. Physiological stress might not always be the same as psychological stress. The predictions are stress probabilities, with near 1 being more stressed.
Device: Garmin Smartwatch

Name: `get_app_usage_blocks`

Description: Retrieves time blocks of app usage for a given user within a specified time range. Each block includes the app name, open time, close time, and duration.

Parameters:

- `uid` (string): The unique identifier for the user.
- `start_time` (string): The start timestamp for the time range.
- `end_time` (string): The end timestamp for the time range.

Returns:
A list of app usage blocks, each containing app name, open and close times, and duration in seconds.

Example:

- {'app': 'SnapChat', 'open': '2024-07-15 17:38:57', 'close': '2024-07-15 18:13:32', 'duration': 2075.0}
- {'app': 'iMessage', 'open': '2024-07-15 19:07:34', 'close': '2024-07-15 19:08:12', 'duration': 38.0}

Fig. 3. Example descriptions of databases (left) and helper functions (right) provided to agents for action plan and code generation.

4 Method

GLOSS is an open-ended query-driven sensemaking system designed to explore and analyze passive sensing data. In this section, we detail our evaluation method for the core GLOSS system using objective and subjective queries. Our aim here is to establish accuracy, consistency, and subjectivity in GLOSS's responses.

4.1 Evaluation Queries

To evaluate GLOSS's performance, we crowd-sourced queries from seven researchers (one Postdoctoral scholar, five Ph.D. students, and one Master's student) working in Ubiquitous Computing research and with prior experience with passive sensing data. We provided researchers with detailed information on the sensor data streams and machine learning models supported by GLOSS. We then instructed the researchers to formulate natural language queries to analyze the data. In the instructions, we asked researchers to draft open-ended queries for a hypothetical placeholder *user ID* and *date range*. Given the running cost and time of the system, we asked researchers to keep the date range up to seven days. We informed the researchers that their queries can ask for triangulating data across multiple datasets, such as, *"How many outgoing calls were made when the user was connected to a wifi on 12/01/25?"*. We include these instructions in our supplementary materials.

This approach resulted in a total of 122 human-generated natural language queries. For each human-generated query, we replaced the placeholder *user ID* with two randomly chosen participants from our testing pool (12 participants, 913 participants-days). Additionally, the `date range` specified in the query was substituted with actual dates where we had data corresponding to the selected participants. For example, the query *"Did userID have an extremely busy day on 26th October 24?"* was transformed into two distinct queries: *"Did u010 have an extremely busy day on 11th June 23?"* and *"Did u013 have an extremely busy day on 25th May 23?"* By applying this approach, we doubled the original query set, resulting in a total of 244 queries, thus allowing us to effectively evaluate the accuracy and reliability of the system. We provide these queries in our supplementary materials.

Next, we followed the query categorizations defined by Merrill et al. [71] and began our coding process by adhering to their definitions of objective and subjective queries:

Objective queries: Objective personal health queries are characterized by clear and specific answers. For example, the question, *"On how many of the last seven days did the user exceed 5,000 steps?"* has a precise and tractable answer.

Subjective queries: Subjective queries are more open-ended and may have multiple valid answers. For example, *“Was yesterday a productive day for the user?”* could yield different responses depending on how productivity is defined.

Two members of the team independently coded the queries, and any disagreements were resolved by consulting a third member of the team. During the coding process, we identified 14 queries that contained both objective and subjective components. Thus, we decided to go back and create a third category where we classified these queries as **mixed queries**. For instance, *"On Jan 5 2025, how many calls did user make while connected to Wi-Fi? Were they multitasking or catching up with someone important?"* has both objective and subjective components. Out of a total of 244 queries, we classified 80.32% (196) as objective, 13.93% (34) as subjective, and 5.73% (14) as mixed queries.

4.2 Evaluation Metrics

Our goal was to compare GLOSS with similar approaches in prior works, but those approaches are close-sourced. Moreover, several of those approaches do not have the capabilities to deal with completely open-ended queries or multi-modal data triangulation. Thus, to compare and evaluate GLOSS, we implemented a baseline model based on a widely used Retrieval-Augmented Generation (RAG) technique in prior works in passive sensing [109, 110]. Similar to prior works, we transformed raw sensor data from different sensor streams into a natural language format [50, 62, 82]. This natural language data is stored in a Chroma database and retrieved based on the user query. We used GPT-4o for the implementation and used LangChain’s RAG framework [3] to implement our baseline RAG model.

Next, we designed separate evaluation metrics for objective and subjective queries. For mixed queries, we evaluated the objective components using objective metrics and the subjective components using subjective metrics. For objective queries, we ran queries in both GLOSS and RAG, setting presentation instructions in GLOSS to *“answer clearly and concisely”*. we evaluated both GLOSS and the RAG in terms of *accuracy* and *consistency*. Accuracy is measured by evaluating whether the model retrieved the correct answer based on the underlying logic it followed. Consistency was measured by passing the same query to the model three times; a response was deemed consistent if it remained identical across all three runs. Some prior works used template-based queries that only covered certain operations and calculations on aggregated data, making the automatic evaluation of accuracy possible. On the other hand, our queries were completely open-ended, requiring the direct use of raw data and the triangulation of data. Thus, to evaluate the accuracy of models, we needed to manually write Python code and analyze the data based on the logic model followed. Our queries spanned over some simple tasks like *“names of all Wi-Fi networks u010 connected on 11th June 23 between 10 am and 7 pm?”* to some complex queries like *“did u010 do any outdoor exercising between 2023-06-11 and 2023-06-17?”* requiring triangulating location, activity, and step data. Thus, we randomly sampled 60 queries [4] from a total of 210 queries (196 objective queries + 14 mixed queries), and the first author wrote code to process & analyze the data to calculate accuracy for these 60 queries. Consistency, however, was evaluated across all 210 queries.

For subjective queries, we generated responses from both GLOSS and RAG, setting presentation instructions in GLOSS to *“explain clearly and in detail”*. We presented the responses to two annotators as *response 1* and *response 2*, ensuring that the source of each response remained hidden. To avoid bias, both *response 1* and *response*

[3] https://python.langchain.com/docs/concepts/rag/

[4] The query dataset size of some prior works [71] is large compared to ours, as their objective queries are designed using templates on aggregated data and thus afford automatic evaluation. In our case, the objective queries were open-ended and required domain experts to manually write code for each query to determine accuracy.

2 included a random number of GLOSS and RAG-generated responses. Following [71], the annotators were tasked to evaluate each model response based on the following attributes: Relevance (relevance of the data used), Interpretation (accuracy in interpreting the question), Domain Knowledge (application of domain knowledge), Logic (logical correctness), and Clarity (clarity of communication). Additionally, they rated the overall reasoning of each response on a Likert scale from 1 ("Poor") to 5 ("Excellent").

5 Results

In this section, we report the performance of GLOSS on both objective and subjective queries, comparing it with a RAG-based baseline (RAG)

5.1 Objective Evaluation

In Table 5, we show the accuracy and consistency comparison between GLOSS and RAG. We performed paired t-tests to measure the significance of accuracy and consistency performances. GLOSS significantly outperforms RAG in terms of accuracy ($t(59) = 8.24, p < .001$). We observe this notable difference due to GLOSS's ability to generate code for calculations and data processing, whereas the RAG-based approach relies solely on LLM calls to interpret the data. This aligns with previous studies indicating that LLMs and RAG-based approaches struggle with mathematical tasks and calculations [3, 59]. The consistency of GLOSS (66.19%) is significantly higher when compared to RAG (52.85%) ($t(59) = 3.33, p = .001$). GLOSS's consistency, however, is lower than its accuracy, primarily due to ambiguity in the queries. Some queries labeled as objective queries by annotators had a single tractable answer provided a particular logic but had multiple valid logical paths. While GLOSS accurately computed the answer, it followed different logical paths in different runs. For instance, for query, *"What day of the week on the first week of June that u010 has the highest mobility?"*, GLOSS used a high step count as an indicator of mobility in one run, while in another, it relied on the maximum difference between traveled GPS coordinates.

Table 5. Comparison of model accuracy and consistency of GLOSS and RAG on objective queries.

Model	*Accuracy %(n)*	*Consistency %(n)*
GLOSS	87.93% (60)	66.19% (210)
RAG	29.31% (60)	52.85% (210)

5.2 Subjective Evaluation

The trends in subjective evaluations were more subtle compared to objective evaluations. In this subsection, we present our results across 5 different criteria. We summarize these trends in Figure 4.

5.2.1 Clarity. To rate *Clarity*, we asked the annotators: *"Is this response easy to read and is it clear?"*. Overall, the annotators found that RAG was able to provide more concise, readable responses compared to GLOSS. However, we found no statistically significant difference between GLOSS and RAG responses ($\chi^2(1) = 1.59, p = .21$). We show the comparison between RAG and GLOSS responses in terms of clarity in Table 6. In the example given in Table 6, GLOSS printed out all timestamps with a heart rate greater than 120 bpm, making the response hard to read and less actionable. It is, however, important to note that RAG demonstrated significantly lower objective accuracy (Table 5), and often led to hallucinated and made-up numbers and inferences that did not actually exist in the data.

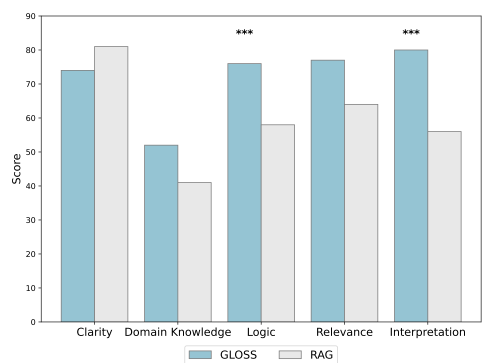

Fig. 4. Comparison of GLOSS and RAG across multiple dimensions. We found statistical evidence between the subjective measurements of GLOSS and RAG in terms of logic and interpretation (highlighted with *** on top of the bars).

Table 6. Comparison between RAG and GLOSS responses in terms of **clarity**. Both models understood the question and were able to come up with a relevant response. Due to its data-driven approach, GLOSS tends to present data in a longer format, while RAG tends to provide a shorter response with *incorrect/made-up data*.

Comparison between GLOSS and RAG	
Query	On June 4, 2023, which Wi-Fi network was u013 connected to when their heart rate went over 120 bpm? Could it have been during an exciting or intense moment?
RAG	On June 4, 2023, u013's heart rate went over 120 bpm at multiple instances: 1. From 12:33:37 onwards, their heart rate reached a peak of 134 bpm, and during this time their phone was connected to a Wi-Fi network labeled "not connected." - *made-up network* 2. Starting from 13:31:39, their heart rate again exceeded 120 bpm and continued to increase. By 13:33:08, their phone was connected to the Wi-Fi network "[anonymized network]." Shortly after at 13:45:39, their heart rate was recorded at lower levels. Given the spikes in heart rate and the records around those time slots, it could suggest that the increases in heart rate were due to some exciting or intense moment, especially looking at the second instance on the "[anonymized network]." network. **Summary:** - u013's heart rate exceeded 120 bpm when connected to the "not connected" network, between 12:44:37 and 12:54:07. - Another spike was recorded while connected to the "[anonymized network]" network around 13:31:39. Without further data, we can reasonably infer that the elevations in heart rate could be related to some form of physical or emotional stress.
GLOSS	On 2023-06-04, user u013 experienced several instances where their heart rate exceeded 120 bpm at the following times: - 10:52:35 - 10:53:05 - 10:53:35 ... (redacted 92 lines) — *overly verbose response* During these periods of elevated heart rate, user u013 was mostly connected to the "[anonymized network]." Wi-Fi network. However, from 12:38:07 to 13:13:08, user u013 was not connected to any Wi-Fi network. Due to an inability to fetch stress level data, it is not possible to definitively determine whether these heart rate spikes were due to exciting or intense moments.

5.2.2 Relevance. Relevance was rated as: *"Does this response reference all relevant user data?"*. Table 7 shows the comparison between RAG and GLOSS responses in terms of relevance. Overall, GLOSS is slightly better at triangulating multiple data sources to answer the query compared to RAG. We found, however, that there is no statistically significance between GLOSS and RAG responses in terms of relevance ($\chi^2(1) = 3.44, p = .06$) In the example shown in Table 7, GLOSS investigated various sensor streams, including step count, location, Wifi, and phone usage pattern, before coming up with the conclusion. In contrast, RAG came to the conclusion after only looking at location and step count data.

5.2.3 Interpretation. Interpretation is rated as *"Does this response properly interpret referenced user data?"*. Table 8 shows an example of the comparison between RAG and GLOSS in terms of interpretation. GLOSS was statistically better at interpreting the data before arriving at a conclusion compared to RAG ($\chi^2(1) = 12.69, p < .001$).

Table 7. Comparison between RAG and GLOSS responses in terms of **relevance**. Both models understood the question and were able to come up with a response. GLOSS, however, searched for evidence using multiple sensor streams before making the final answer, while RAG answered using only two sensor streams - location and step count.

Comparison between GLOSS and RAG	
Query	Will test007 be more likely to step out of the house on 10/16/24 based on past two days of data?
RAG	Based on the provided sensor data for the past two days, it appears that test007 has been stationary and has not recorded any steps or movement. On both October 14th and October 15th, the data shows that test007 remained in the same locations without any significant activity. Given this pattern of inactivity, it is unlikely that test007 will step out of the house on October 16, 2024.
GLOSS	Based on the data collected for user_id test007 for the dates 2024-10-14 and 2024-10-15, we observe the following insights: **Activity Data**: The user was primarily stationary with periodic walking intervals. For example, on 2024-10-14, there was notable stationary time from 12:21 AM to 9:36 AM with brief walking episodes. **Location Metrics**: The user visited 10 different significant locations with substantial displacements. The maximum displacement recorded was 18,758.52 meters, and the total travel distance was 36,829.64 meters over the two days. **Steps Data from Phone**: The user took a total of 3,844 steps, covering 2,783.15 meters, and ascended 10 floors while descending 9 floors. **Wi-Fi Connection Logs**: The user connected and disconnected from multiple Wi-Fi networks, such as [anonymized networks], indicating movement and not being in a single location for extended periods. **Phone Lock/Unlock Times**: The phone was locked 123 times and unlocked 124 times, with the phone being unlocked for a total of approximately 19 hours and 21 minutes. This comprehensive data, showing consistent activity levels, movement patterns, Wi-Fi connections, and phone usage, suggests that the user had a pattern of stepping out over the past two days. Therefore, it is likely that user test007 will step out of the house on 2024-10-16. *— triangulating multiple database to draw conclusions*

In the example shown in Table 8, both RAG and GLOSS understood the question and provided evidence for the responses based on data. RAG, however, made a false interpretation of the data at the end of the response. RAG inferred "not connected" as a wifi network, whereas it meant the phone was not connected to any wifi network at that time.

5.2.4 Domain Knowledge. Table 9 shows an example of the comparison between RAG and GLOSS in terms of domain knowledge. Domain Knowledge is rated as *"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)?"*. Overall, GLOSS was slightly better at interpreting the data before arriving at a conclusion compared to RAG, especially in query that requires the use of common sense knowledge. We found, however, that there is no significance between GLOSS and RAG responses in terms of domain knowledge ($\chi^2(1) = 1.9, p = .17$).

In the example shown in Table 9, both models understood the question and referenced relevant data to answer. However, RAG failed to detect the abnormality in the Garmin step count data and reached the wrong conclusion.

Table 8. Comparison between RAG and GLOSS responses in terms of **interpretation**.

Comparison between GLOSS and RAG	
Query	On Aug 1 2024, how many steps did test006 take while connected to wi-fi? it might show moments of movement at home or work.
RAG	On August 1, 2024, while connected to Wi-Fi, test006 took the following steps: **From 08:42:46 to 10:37:53 (connected to <institution wifi name>):** - 83 steps between 08:49:59 and 08:51:06. **From 20:48:02 to 21:01:02 (connected to not connected):** wrong inferencing thinks "not connected" is a wifi name - 298 steps between 20:49:11 and 20:52:17. - 38 steps between 20:53:13 and 21:04:45 (partially overlapping with the next Wi-Fi connection window). Adding these up: 83 steps (<institution wifi name>) + 298 steps (not connected) + 38 steps (partially in the not connected window) = 419 steps. Therefore, test006 took **419** steps while connected to Wi-Fi on August 1, 2024.
GLOSS	On August 1, 2024, test006 took a total of **9325** steps while connected to Wi-Fi. Here are the details: **Connected to <institution wifi name>:** - From 05:05:19 to 08:31:46, took 1079 steps - From 08:42:46 to 10:37:53, took 1280 steps <redacted> **Connected to 'xfinitywifi':** - From 16:00:25 to 17:03:58, took 581 steps <redacted> In total, test006 took **9325** steps while connected to Wi-Fi on 2024-08-01.

In contrast, GLOSS not only identified the discrepancy between the phone and Garmin step count but also explained it using domain knowledge. This highlighted GLOSS' ability to incorporate data from multiple sensors and arrive at the correct conclusion.

5.2.5 Logic. Table 10 shows an example of the comparison between RAG and GLOSS in terms of logic. Logic is rated as *"Does this response contain evidence of correct reasoning and logic?"*. Overall, GLOSS was statistically better at reasoning and logical conclusion compared to RAG, especially in queries that require calculation of quantitative data ($\chi^2(1) = 6.64, p = .001$).

In the example shown in Table 10, both models understood the question, referencing relevant data, and interpreting the data correctly. RAG, however, made a calculation mistake in its response, possibly due to hallucination. GLOSS, on the other hand, never uses LLM for calculation but relies on code generation, which reduces the likelihood of making calculations and logical mistakes.

6 Potential Use Cases of GLOSS

GLOSS provides a powerful system for analyzing and understanding passive sensing data, transforming what is often considered a complex and time-consuming process into a more accessible and efficient one. Its open-ended sensemaking capabilities make GLOSS a promising solution for various applications in personal health informatics (PHI) and well-being monitoring. In this section, we present multiple potential use cases of GLOSS, drawing

Table 9. Comparison between RAG and GLOSS responses in terms of **domain knowledge**. In this example, RAG was unable to notice the abnormality in Garmin step count data, and arrived at the wrong conclusion. GLOSS, however, was able to explain the discrepancy between the phone and Garmin step count using common knowledge.

Comparison between GLOSS and RAG	
Query	Did test008 a physically active person in general based on data from 06/25/24 to 06/29/24?
RAG	Based on the data provided from 06/25/24 to 06/29/24, test008 does not appear to be a physically active person during this period. The data shows that test008 consistently walked 0.0 steps, covered 0.0 meters, and remained stationary for multiple periods each day. There is no record of physical activity in the form of steps taken or floors climbed or descended. Therefore, it can be concluded that test008 was not physically active during the specified dates.
GLOSS	Based on the data from June 25, 2024, to June 29, 2024, user test008 appears to be generally physically active. **Activity Data:** - Stationary: 101 times for a total of 4 days, 12 hours, and 49 minutes - Walking: 72 times for a total of 5 hours and 42 minutes - Automotive: 65 times for a total of 1 hour and 52 minutes - Running: 10 times for a total of 15 minutes - Cycling: 59 times for a total of 2 hours and 37 minutes **Phone Steps Data**: - Total Steps: 32,688 - Average Steps per Day: 6,537.6 **Garmin Steps Data**: - Total Steps: 0 - Average Steps per Day: 0 Despite the zero steps recorded on the Garmin device (likely due to not wearing the device or data retrieval issues), the variety and duration of physical activities logged and phone steps data indicate that test008 was engaging in physical activities and was generally physically active during this timeframe. *— ability to give explanation for abnormal data*

inspiration from prior and ongoing work in UbiComp and the HCI communities. While the scope of this paper focuses on the core GLOSS system with demonstrations of potential use cases, we plan to conduct a thorough evaluation of each of these use cases as part of our future work.

To make a user-friendly way to interact with GLOSS, we implemented *a chat-based user interface (UI)* for GLOSS using the Streamlit Python library. The GLOSS UI displays different components of sensemaking like *action plan*, *information requests*, *memory*, and *understanding* updating them as the iterative sensemaking process continues. We present the GLOSS UI in Figure 5. The user begins by entering a query and presentation instructions into the search box at the top of the screen (A). The sensemaking process starts when the user presses "Start Sense-Making" button. Throughout the process, the status bar (B) shows the current state of the sensemaking process (e.g., information seeking, local sensemaking, code generation, etc.) As the process continues, different components of sensemaking (*action plan*, *information requests*, *memory*, and *understanding*) are updated (C). Finally, when the sensemaking process finishes, the final answer is presented to the user, tailored according to their presentation instructions (D).

Table 10. Comparison between RAG and GLOSS responses in terms of **logic**. In this example, RAG was making a minor calculation mistake at the end of its response.

Comparison between GLOSS and RAG	
Query	Does test006 prefer texting over calling based on data from 06/08/24 to 06/13/24?
RAG	Based on the data provided from 06/08/24 to 06/13/24: **Phone call durations:** - 06/08/24: 3 seconds, 29120 seconds (approximately 8.09 hours) - 06/09/24: 5 seconds, 532 seconds (approximately 8.87 minutes), 2 seconds, 45 seconds, 3 seconds - 06/11/24: 2 seconds, 293 seconds (approximately 4.88 minutes) **iMessage durations:** - 06/08/24: 3 seconds, 0 seconds, 0 seconds, 9 seconds - 06/09/24: 5 seconds, 19 seconds, 9 seconds, 0 seconds, 60 seconds, 0 seconds, 9 seconds, 10 seconds - 06/11/24: 102 seconds (approximately 1.7 minutes), 19 seconds, 113 seconds (approximately 1.88 minutes), 7 seconds **Summarizing the total duration spent on phone calls versus iMessage:** - Total phone call time: 29120 + 5 + 532 + 2 + 45 + 3 + 2 + 293 = 29902 seconds (approximately 8.3 hours) — *wrong calculation, should be 30,002 seconds. The overall logic remains wrong even if this calculation was right.* - Total iMessage time: 3 + 0 + 0 + 9 + 5 + 19 + 9 + 0 + 60 + 0 + 9 + 10 + 102 + 19 + 113 + 7 = 365 seconds (approximately 6.08 minutes) Considering the significant difference in time spent, it is evident that test006 prefers calling over texting. The total duration spent on phone calls far exceeds the time spent using iMessage.
GLOSS	Based on the data available, for user_id 'test006' between 2024-06-08 and 2024-06-13, the total duration of usage for messaging apps is 12 hours, 51 minutes, and 33 seconds, while the total duration of calls made and received is 37 hours, 37 minutes, and 6 seconds. This indicates that during this period, test006 spent significantly more time on phone calls than on texting. Therefore, test006 prefers calling over texting during the specified date range.

6.1 Supporting Non-computer Science Researchers with Sensemaking

With the ubiquitous presence of mobile phones and wearables in people's lives, behavioral scientists, clinicians, and psychologists have shown a growing interest in passive sensing data to study people's behavioral patterns [4, 84]. Passive sensing has the potential to unobtrusively collect and reveal meaningful affective and behavioral patterns over time [1, 76]. These researchers, although interested in passive sensing data, often lack computer science training and experience (we call them non-CS researchers). This lack of expertise creates substantial barriers for non-CS researchers in independently processing, analyzing, and interpreting longitudinal passive sensing data [36]. As a result, these researchers often depend on the expertise of trained computer scientists and data analysts to process and interpret passive sensing data [107]. While this collaboration is valuable, it can also lead to challenges. For instance, if non-CS researchers require additional data or specific analyses, they must request assistance from their computer science collaborators and/or staff. This process can often be time-consuming, leading to considerable delays and increased effort for everyone.

GLOSS provides a promising solution to this barrier faced by non-CS researchers. Its chat interface empowers them to interpret and analyze passive sensing data independently. GLOSS can generate code, perform data triangulation, and present insights in a transparent and interpretable format customized to their specific needs. To provide a concrete example of GLOSS's potential impact, we conducted semi-structured interviews and a

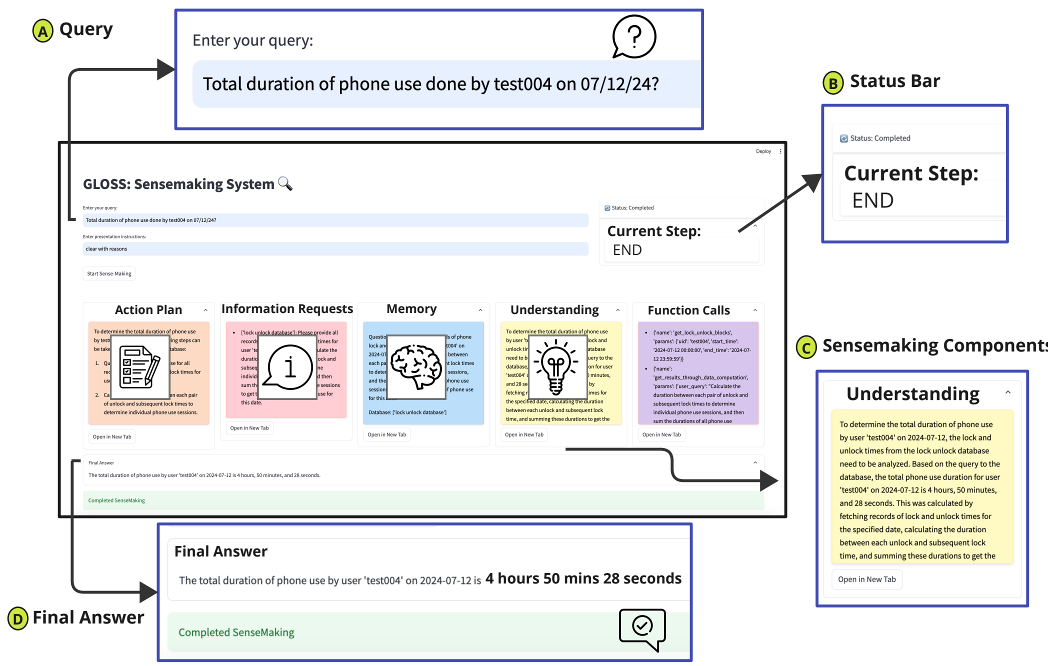

Fig. 5. GLOSS UI design that shows relevant components of sensemaking (such as Action Plan, Information Requests, etc.). As the sensemaking process continues, these different components in the UI get updated with relevant information.

think-aloud session with three psychologists from a research group working on passive sensing data to understand adolescents' mental health and well-being at a private R1 institution. We set up GLOSS with a subset of their study data so they were acutely familiar with their dataset and could have an effective interaction with GLOSS. Each interview session lasted for 30-45 minutes. We started the interview by asking questions about their current practices in dealing with passive sensing data, followed by a tutorial demonstrating how to use GLOSS through its user interface. Next, in a think-aloud session, participants interacted with GLOSS, which lasted for around 10-15 minutes. We asked participants to come up with open-ended queries based on data monitoring and analysis tasks they encounter on a day-to-day basis. After the session, we asked participants questions about their experience and the perceived usefulness of GLOSS. At the end of the interview, participants completed a modified version of the System Usability Scale (SUS). This study was approved by our IRB, and participants were compensated $15 for their time.

We conducted the meetings in person and recorded the audio and screen interaction using Zoom. The audio from the meetings was automatically transcribed using the Microsoft Word transcriber. Two members of our research team independently coded the interviews using the Open Coding approach [22]. The coders later met to coalesce common themes and discuss disagreements. We present the main themes that emerged from the interview sessions below:

6.1.1 Current analysis and monitoring methods: All the participants relied on their existing study dashboards for data visualization and monitoring. P3 explained, *"Right now our team has a dashboard that allows us to see different aspects of different data streams we're collecting."* Our participants did not have formal training in computer science (mean self-reported coding proficiency was 3.34/10). Some participants mentioned that they use R for statistics and calculation sometimes, but would not consider themselves proficient. Due to a lack of technical expertise, participants mentioned that accessing data not available on the dashboard requires reaching out to technical team members. They described this process as often time-consuming and requiring considerable effort. As P2 said, *"I would have to reach out to the data scientists and computer scientists that our team works with and ask them for support (for getting additional data) because I'm not necessarily sure what the best way to go about that would be"*. Even after receiving data from technical staff, participants mentioned that they often need to manually review and, at times, hand-code data, which is a highly labor-intensive process. P1 elaborated on this, sharing, *"Yes (it is time-consuming), especially when I'm hand coding when I know it would be a lot easier not to hand code, but it feels easier than having to learn how to navigate some CS system. It is scary."*. Participants expressed that this manual coding method would not be scalable if they had a large participant pool in their studies.

6.1.2 Experience with GLOSS:. Participants found GLOSS to be a valuable tool for monitoring and analyzing data, highlighting its ability to save significant amounts of time. P2 shared, "*Overall, it was a really amazing experience. I can already see it saving so much time. Maybe a minute from typing this to seeing the result [...], and it's absolutely miraculous that we can do it this quickly.*" They also felt that GLOSS made complex tasks much more manageable. GLOSS was described as easy to use, with P1 noting, "*Since you're able to just kind of ask the question however it pops up in your head, you don't really have to think through the exact code or anything to figure it out. It makes it much more accessible.*" Similarly, P2 remarked, "*You can use it (GLOSS) like a Google search. People are very comfortable with that.*"

Participants also appreciated GLOSS's user interface, which displayed various components such as the *action plan*, *information requests*, *memory*, and *understanding*. They felt that showcasing how GLOSS performs sensemaking through these elements improved the system's explainability and strengthened their trust in it. P2 expressed, *"It's nice that it's (GLOSS) just so transparent because I feel like that's a big issue that people talk about with these things, with language learning models. You'll get this output, but then it's like, OK, how did we get here? Can we really trust it? For things like research, when you need to be so precise and accurate, it's really helpful that we can see each step it went through.*" The SUS survey reaffirmed that participants viewed GLOSS as a valuable, easy-to-use, and easy-to-learn tool, with an average aggregated SUS score of 84.02%

6.1.3 Improvements to GLOSS:. We asked participants for suggestions on improving GLOSS to better meet their needs. One common idea was to allow for query history and follow-up questions. P3 mentioned, "*If the response I got didn't align exactly with what I was looking for, it'd be cool to follow up on it.*" Meanwhile, P2 suggested, "*Maybe you could add a way to toggle through previous queries to see what those were.*" Participants also felt that having instructions on how to prompt GLOSS for the best results would be helpful. P2 expressed, "*The only thing I worry about is whether I have the correct phrasing for the query so the model knows what I'm trying to ask. A cheat sheet for prompting, especially for people less familiar with language models, would be useful.*" This reflects recent research showing that non-experts struggle with prompt design [111]. Other suggested improvements included the ability to download code and data directly from the GLOSS interface. P1 also mentioned that having a feature to generate graphs in GLOSS would be very useful.

These interviews serve as a strong indicator of GLOSS's potential impact in assisting sensemaking for non-CS researchers. In the future, we plan to conduct more comprehensive evaluations by deploying GLOSS in multiple research labs that work with passive sensing data.

6.2 Narratives for Personal Health Reflection

Reflection plays an important role in the journey of personal health informatics [60]. It enables individuals to make meaningful connections with their health data [61] inspiring positive lifestyle changes [12, 31] and goal-setting behaviors [38, 78]. Merely presenting data, however, might not trigger reflection [9]. Moreover, traditional ways of presenting data using tables and graphs often reduce individuals to quantified beings [10] overlooking more qualitative and holistic views of wellbeing. More recently, researchers used LLMs to generate personal health narratives that resonate more deeply with individuals' experiences and perceptions of wellbeing [35, 71, 96]. Stromel et al. [96] demonstrated that supplementing step count visualizations with narrative descriptions significantly enhanced users' reflective engagement with their personal health data. GLOSS can be used to take this approach further by triangulating data from multiple streams and synthesizing multi-dimensional personalized narratives to support reflective practices. Furthermore, through presentation instructions, GLOSS can tailor data narratives to be adjusted to an individual's needs. In Figure 6, we show that using the same query but with different presentation instructions, GLOSS can generate narratives specifically tailored to different aspects of well-being (qualitative vs quantitative).

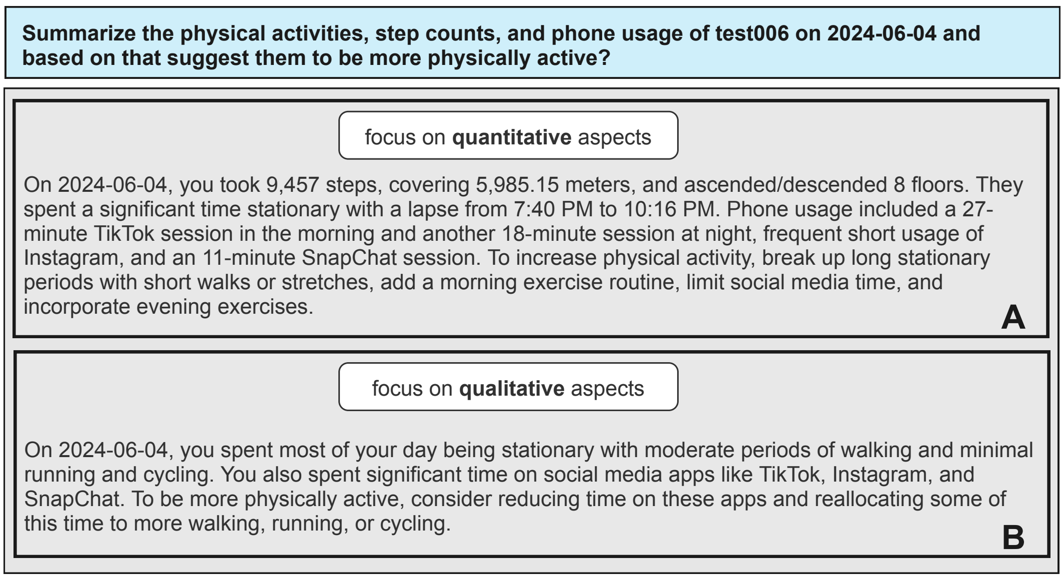

Fig. 6. GLOSS generates different narratives to the same query when instructed to focus on quantitative and qualitative dimensions of wellbeing.

6.3 AI-triggered Prompting

Ecological momentary assessments (EMA) [94] have been used extensively to collect various subjective measures of health and wellbeing, such as, depression [101, 108], anxiety [98], daily activities [55, 56], and others [13]. Traditionally, EMA prompts have been delivered at fixed times during the day. Recent works, however, have leveraged AI models to identify opportune moments for delivering these prompts, a method known as AI-triggered prompting [41, 64, 73]. For instance, researchers have used AI models to predict physiological stress and use it as a basis for triggering EMA prompts aimed at capturing stressors [77]. Understanding the nature of stressors can help researchers design tailored and effective interventions.

While EMAs offer valuable insights into perceived subjective well-being, participants often find them burdensome [13, 20, 43], with the level of burden increasing as the length of the questionnaire increases [100]. We demonstrate a use case for GLOSS where it can reduce the length of EMAs by only prompting contextually relevant questions. GLOSS can run in the background and keep integrating data from multiple sensor streams to continuously analyze and understand an individual's context. This context-aware understanding can be used to trigger only the prompts that are relevant to the context. For instance, in Figure 7 B, due to a change in location and connected Wi-Fi networks, GLOSS posed a question asking whether the change in surroundings was a potential cause of stress. Moreover, it can also help in addressing missing data in passive sensing through relevant questions. For example, in Figure 7 A, due to the absence of app usage and phone conversation data—likely caused by technical issues—GLOSS proactively poses a question to address this data gap.

6.4 Enhancing Interpretability

Deep learning models are extensively used to detect anomalous social media and phone use and its association with various mental health outcomes [15, 25, 105]. Understanding these anomalous patterns can provide researchers with insights into how social media and phone use relate to various affective and mental health outcomes, ultimately aiding in the development of effective interventions [103]. While these anomaly detection models demonstrate high accuracy, they are often considered black boxes, lacking interpretability [113]. This limitation can hinder the ability to derive actionable steps or interventions [32, 49]. To address this issue, we present an example of how GLOSS can be integrated with traditional models to enhance interpretability and support researchers in taking actionable steps. In an ongoing project focused on detecting anomalous app usage, we implemented an unsupervised outlier detection model based on Long Short-Term Memory (LSTM) networks. These personalized models analyzed the past seven days of data to identify whether app usage on a given day was anomalous. As LSTMs are deep learning-based models, their predictions lacked interpretability. To overcome this, we augmented model predictions with insights derived from GLOSS (Figure 8), enhancing the interpretability of the results. Researchers can leverage these insights as a starting point and can query GLOSS for additional information to formulate actionable strategies.

7 Discussion

In this section, we discuss the results from our evaluations of GLOSS as well as some implications and limitations of our work.

7.1 GLOSS: Enabling Open-ended Sensemaking

Researchers in the past have utilized passive sensing data to develop prediction and detection models for multiple but specific behavioral and health outcomes such as depression, stress, and physical activity [25, 75, 98, 106]. While these detection systems have shown tremendous potential, they often reduce complex aspects of health and well-being to binary predictions, failing to capture the broader, holistic picture [10]. There is a growing call within the research community to move beyond quantitative assessments and develop a more comprehensive understanding of well-being [2, 34, 96]. To achieve this, we need systems that make passive sensing data more accessible and interpretable to a diverse range of stakeholders, including researchers, behavior scientists, clinicians, and even self-health trackers themselves. To this end, we designed a sensemaking system GLOSS (**G**roup of **L**LMs for **O**pen-ended **S**en**s**emaking) inspired by prior theories on human sensemaking [51, 83]. GLOSS is a query-based system that works with high-dimensional raw passive sensing data. GLOSS uses code generation techniques that are capable of writing and executing code to perform advanced data processing and analysis. Furthermore, GLOSS is able to triangulate multiple data streams to derive insights and understand the context of individuals. Using GLOSS does not demand any expertise in passive sensing or programming. Overall, GLOSS enables open-ended sensemaking, which was not possible with existing systems, paving the way for new applications and research directions.

User u011 reported high stress around 8.30 pm on 2023-05-28. Look at all the information you can to understand the context around that time? Present the context and use the context to generate three question to understand the reason of stress better.

Context:
On 2023-05-28 around 8:30 PM, user u011 reported high stress. From the Garmin stress database, the highest stress probability was recorded as 0.49 at 8:31 PM. Heart rate data between 8:00 PM and 9:00 PM showed fluctuations between 66 bpm and 125 bpm. The activity database indicated that the user was walking, stationary, and running around 8:27 PM. Location data revealed multiple movements within a localized area with coordinates including (latitude1, longitude1), (latitude2, longitude2), (latitude3, longitude3), and others. There were no app usage or phone interactions recorded during this timeframe. User u011 walked 1880 steps according to the phone steps database and 2253 steps according to the Garmin steps database, covering a distance of 1192.08 meters, with no floors ascended or descended.

Questions:
- Were you engaged in any physical activity before or around the time you felt stressed?
- Were you using any specific apps or interacting with your phone in a way that might have contributed to your stress?
- Did you have any phone conversations or were you at any particular location that might have caused you stress during that time?

A

User test006 reported high stress around 7.00 pm on 2024-06-04. Look at all the information you can to understand the context around that time? Present the context and use the context to generate three question to understand the reason of stress better.

Context:
The Garmin stress database confirms that user test006 experienced elevated stress levels between 18:47:45 and 18:53:45, with stress probabilities peaking at 0.81 at 18:47:45. According to the activity database, the user was mostly stationary from 18:00 to 19:00, with brief periods of walking between 19:06 and 19:14, cycling from 19:13 to 19:19, and walking again from 19:18 to 19:29. The location data shows that user test006 remained in an area near (latitude1, longitude1), before 19:00, moving to another area localized around (latitude2, longitude2),during the elevated stress period. Additionally, data from the wifi database indicates a change in wifi networks around 19:03:14, and the user engaged with various apps such as Instagram, TikTok, and Camera between 18:00 and 20:00. The call log database shows no phone calls made or received within that timeframe.

Questions:
- Were you involved in any challenging or stressful activities (e.g., work, meetings, exercise) around 19:00 on 2024-06-04?
- Did your location or surroundings change significantly around 19:00, and if so, how do you feel about those places?
- Were there any specific app activities that could have contributed to your stress levels around that time?

B

Fig. 7. GLOSS can model the context using passive sensing data and then use the context to generate relevant questions to understand the stressor.

7.2 Comparison of GLOSS and RAG Performance

Through our evaluations, we establish that GLOSS has promising accuracy (Table 5), which is significantly higher than the commonly employed RAG techniques [110]. Our results align with prior research that shows RAG techniques perform poorly on mathematical or computational tasks [3, 59]. The consistency of GLOSS was affected by the ambiguity of queries, leading to GLOSS choosing accurate but different logical steps in

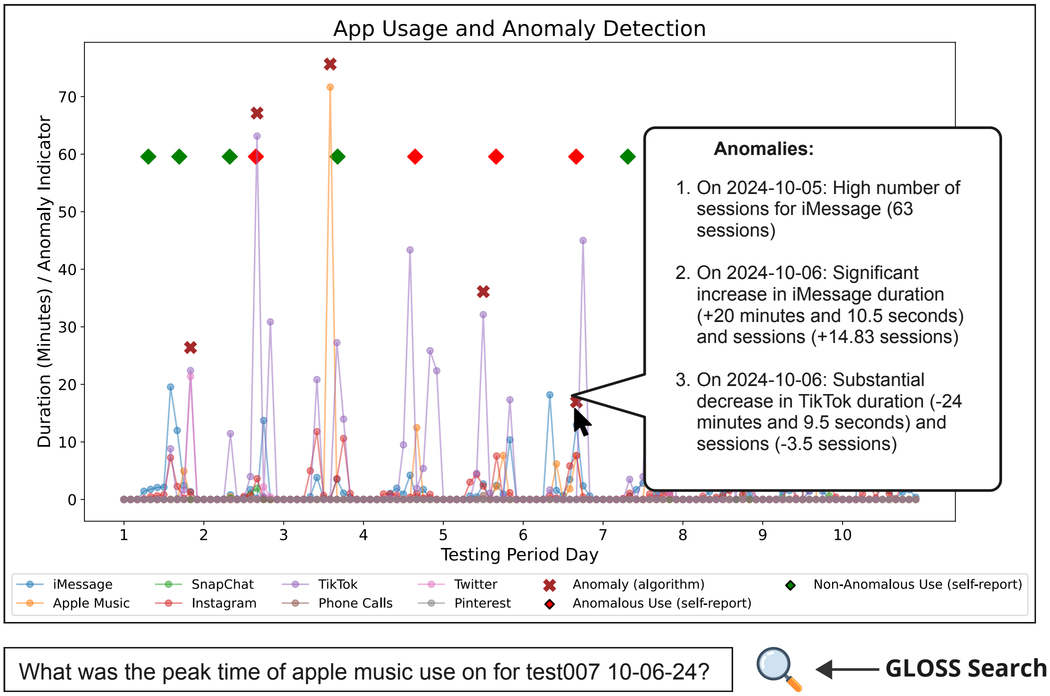

Fig. 8. Black box models to be augmented with GLOSS-generated insights to enhance interpretability of results supporting actionable steps for researchers.

different runs. For instance, when answering the query "How many times did test008 use their phone during the commute on 2024-07-12?", GLOSS took different logical paths in separate runs: in one instance, it identified commute periods based on changes in location, while in another, it relied on activity recognition (walking or automotive) to determine commute times. While different logical steps can encourage stakeholders to consider multiple perspectives on ambiguous queries, they can also create confusion. To mitigate this, we plan to introduce a preliminary step before sensemaking starts, where GLOSS can request more information to make the query more specific if it is deemed ambiguous. In our example, we can ask the user to clearly define *commute* or ask the user to choose one from possible options to calculate commute times (e.g., using *activities* or using *location*) to resolve the ambiguity. Another approach would be to expand GLOSS further to be comprehensive and combine several valid approaches to deal with potentially ambiguous tasks.

In our subjective evaluations, GLOSS responses outperformed the RAG baseline in relevance, interpretation, domain knowledge, and logic, while RAG was rated higher in clarity. Based on the reasonings of our annotators, we believe GLOSS was over-explaining details as we set presentation instructions to *"Explain clearly and in details"*. We did not do extensive experiments on presentation instructions affecting subjective evaluations of GLOSS responses. Although GLOSS outperformed RAG in most subjective dimensions, these differences were subtle, which is concerning, as RAG responses were highly inaccurate. Our results are supported by prior works that have shown that LLMs can generate believable, confident and clear explanations for inaccurate results [26, 53, 99]. These misleading explanations can be potentially detrimental, especially if applied in the health and well-being domain (e.g., for self-reflection, intervention design, etc.). Thus, researchers should be cautious when using RAG for passive sensing tasks involving computation and data triangulation.

7.3 Potential for Future Applications

We envision GLOSS as a foundational model for open-ended sensemaking, capable of being used directly or built upon for multiple use cases, including in human-facing or automated applications. In Section 6, we demonstrated four applications of GLOSS inspired by ongoing and prior research in passive sensing. First, we conducted semi-structured interviews and a think-aloud session with three psychologists working with passive sensing data on a daily basis. The psychologists found GLOSS to be a highly valuable, time-saving, transparent, and easy-to-use system. They expressed that GLOSS enables them to perform sensemaking and analysis that was previously beyond their expertise. Second, we showed that GLOSS can be used to generate multi-dimensional narratives of personal health, which can be used to trigger reflective practices. Third, we showed GLOSS's application in generating relevant EMA prompts based on the context of an individual. Fourth, we showed how GLOSS can add to the interpretability of black box models, which can help in deriving actionable insights. These applications covered a diverse range of stakeholders, including psychologists (non-CS researchers), intervention designers, self-health trackers, and researchers. Through these examples, we wanted to highlight the breadth of applications of GLOSS, keeping a thorough evaluation of these use cases as part of our future work. As GLOSS is a core component that can be integrated with many human-facing and backend systems, GLOSS has several other potential applications. For instance, researchers can integrate GLOSS with text-to-speech and speech-to-text systems to enhance the accessibility of personal health data to individuals with vision-related disabilities and older adults who find visualization-only interfaces inaccessible. Furthermore, as the design of GLOSS does not restrict it to passive sensing, it can be extended for sensemaking in other forms of data (e.g., images and audio). Overall, we believe that GLOSS is a valuable sensemaking system for the UbiComp and HCI communities, and by making our code and results publicly available, we hope to support other researchers in exploring these possibilities with GLOSS.

7.4 Future Improvements to GLOSS

Based on our evaluations, we identified opportunities for improvements that can further enhance sensemaking using GLOSS. Some of these improvements were highlighted by participants in Section 6.1.3. First, GLOSS is currently a single-turn query-based system. It does not keep a history of previous queries, nor does it allow users to ask follow-up questions. This hinders users from iteratively using GLOSS to understand data, building on previous GLOSS responses. In the future, we plan to enable multi-turn conversations and keep a history of queries and results. Second, GLOSS UI currently does not support showing graphs or downloading data and code; we can add these functionalities to enhance the interpretability and transparency of results and broaden the scope of use for researchers interested in using GLOSS for their studies. Third, GLOSS is sensitive to ambiguous queries, and some non-CS researchers have expressed concern about whether they are providing the correct prompts to the model. As non-experts often face trouble with prompt designing [111], we can include instructions on prompting with some examples in GLOSS UI to help users design effective prompts. Furthermore, with multi-turn conversation, GLOSS could request the user to clarify the query further in case it was unclear. These improvements can make GLOSS an even more powerful and valuable system.

7.5 Limitations

While our work offers a novel and valuable system for open-ended sensemaking, it has some limitations. First, we did not experiment with using multiple LLMs. As different LLMs have different code generation capabilities, our results may vary across different LLMs. Second, as GLOSS calls the GPT-4o model multiple times, its latency depends on the latency of these API calls and network connectivity. Thus, sometimes GLOSS has a latency of a couple of minutes when it processes complex queries, which can be irritating to users in applications where they expect near real-time latency. Having a local LLM in the system running GLOSS can reduce latency significantly. Third, we did not experiment with modifying presentation instructions and its impact on the subjective evaluation of user queries. Modifying presentation instructions can change the answer and thus their subjective evaluation on dimensions like *clarity* and *interpretation*. Fourth, we used passive sensing data modalities from phones and wearables. While GLOSS functioning does not rely on a single type of data, we still might need more evaluations to establish GLOSS's performance on other kinds of passive sensing data (for instance, ambient temperature sensors, audio devices, etc.). Lastly, our query dataset for evaluation was small and was generated by researchers working in the passive sensing field. Their queries naturally leaned more toward the objective inquiries they often encountered in their research, leading to more objective queries than subjective queries in our evaluation dataset. As future work, we will evaluate the effectiveness of GLOSS in diverse populations and query distributions.

8 Conclusion

In the work, we presented a query-based novel sensemaking system, GLOSS, where LLM agents interact according to theories on human sensemaking. GLOSS fills the gaps in existing sensemaking systems by enabling accurate, open-ended, and multimodal understanding of passive sensing data. We presented four use cases of GLOSS for automated tasks and human-facing tasks. The human subjects highly liked GLOSS, deeming it a valuable system that can save significant time and effort. We firmly believe GLOSS can be used to drive several interesting future health and well-being applications.

Acknowledgments

This research is partially supported by the National Institutes of Health, under award number NIDA P30DA029926, and the National Science Foundation, under award number IIS-2442593, and compute resources provided by Microsoft's AICE (AI, Cognition, and the Economy). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors. Any mention of specific companies or products does not imply any endorsement by the authors, by their employers, or by the sponsors.

References

[1] Saeed Abdullah and Tanzeem Choudhury. 2018. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia* 25, 1 (2018), 61–75. Publisher: IEEE.

[2] Daniel A. Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C. Mohr, Anna R. Van Meter, Julia C. Tartaglia, Nicholas C. Jacobson, Fei Wang, Deborah Estrin, and Tanzeem Choudhury. 2024. Beyond detection: Towards actionable sensing research in clinical mental healthcare. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (Nov. 2024). https://doi.org/10.1145/3699755 Number of pages: 33 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 160 tex.issue_date: December 2024.

[3] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation. https://doi.org/10.48550/arXiv.2407.12216 arXiv:2407.12216 [cs].

[4] Chioma Virginia Anikwe, Henry Friday Nweke, Anayo Chukwu Ikegwu, Chukwunonso Adolphus Egwuonwu, Fergus Uchenna Onu, Uzoma Rita Alo, and Ying Wah Teh. 2022. Mobile and wearable sensors for data-driven health monitoring system: State-of-the-art and future prospect. *Expert Systems with Applications* 202 (2022), 117362. Publisher: Elsevier.

[5] Neeraj K Arora, Bradford W Hesse, Barbara K Rimer, Kasisomayajula Viswanath, Marla L Clayman, and Robert T Croyle. 2008. Frustrated and confused: the American public rates its cancer-related information-seeking experiences. *Journal of general internal medicine* 23 (2008), 223–228. Publisher: Springer.

[6] Christiane Attig and Thomas Franke. 2020. Abandonment of personal quantification: A review and empirical study investigating reasons for wearable activity tracking attrition. *Computers in Human Behavior* 102 (2020), 223–237. Publisher: Elsevier.

[7] Ling Bao and Stephen S. Intille. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Alois Ferscha, and Friedemann Mattern (Eds.). Vol. 3001. Springer Berlin Heidelberg, 1–17. https://doi.org/10.1007/978-3-540-24646-6_1 Series Title: Lecture Notes in Computer Science.

[8] Emma Caroline Barrett. 2009. *The interpretation and exploitation of information in criminal investigations*. phd. University of Birmingham.

[9] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*. 93–102.

[10] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27. Publisher: ACM New York, NY, USA.

[11] Laura Bloomfield, Mikaela Fudoligb, Peter Doddsb, and Llorinb Jordan. [n. d.]. Detecting stress in college freshman from wearable sleep data.

[12] Fadi Botros, Charles Perin, Bon Adriel Aseniero, and Sheelagh Carpendale. 2016. Go and grow: Mapping personal data to a living plant. In *Proceedings of the international working conference on advanced visual interfaces*. 112–119.

[13] Cynthia I Campbell, Ching-Hua Chen, Sara R Adams, Asma Asyyed, Ninad R Athale, Monique B Does, Saeed Hassanpour, Emily Hichborn, Melanie Jackson-Morris, Nicholas C Jacobson, and others. 2023. Patient engagement in a multimodal digital phenotyping study of opioid use disorder. *Journal of Medical Internet Research* 25 (2023), e45556. Publisher: JMIR Publications Toronto, Canada.

[14] Stefano Canali, Viola Schiaffonati, and Andrea Aliverti. 2022. Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. *PLOS Digital Health* 1, 10 (2022), e0000104. Publisher: Public Library of Science San Francisco, CA USA.

[15] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1 (2020), 43. Publisher: Nature Publishing Group UK London.

[16] Wenqiang Chen, Jiaxuan Cheng, Leyao Wang, Wei Zhao, and Wojciech Matusik. 2024. Sensor2Text: Enabling natural language interactions for daily activity tracking using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–26. Publisher: ACM New York, NY, USA.

[17] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E Palacios, Derek Richards, and Anja Thieme. 2020. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[18] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1143–1152.

[19] Akshat Choube, Rahul Majethia, Sohini Bhattacharya, Vedant Das Swain, Jiachen Li, and Varun Mishra. 2024. Imputation Matters: A Deeper Look into an Overlooked Step in Longitudinal Health and Behavior Sensing Research. https://doi.org/10.48550/arXiv.2412.06018 arXiv:2412.06018 [stat].

[20] Akshat Choube, Vedant Das Swain, and Varun Mishra. 2024. SeSaMe: a framework to simulate self-reported ground truth for mental health sensing studies. *arXiv preprint arXiv:2403.17219* (2024).

[21] Martina A Clarke, Joi L Moore, Linsey M Steege, Richelle J Koopman, Jeffery L Belden, Shannon M Canfield, Susan E Meadows, Susan G Elliott, and Min Soon Kim. 2016. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: A literature review. *Health informatics journal* 22, 4 (2016), 992–1016. Publisher: SAGE Publications Sage UK: London, England.

[22] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21. Publisher: Springer.

[23] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A. Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, Megan Walker, Jeffrey Yu, Tim Althoff, Conor Heneghan, John Hernandez, Mark Malhotra, Leor Stern, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Jiening Zhan, Shruthi Prabhakara, Daniel McDuff, and Cory Y. McLean. 2024. Towards a Personal Health Large Language Model. https://doi.org/10.48550/arXiv.2406.06474 arXiv:2406.06474 [cs].

[24] Aykut Coşkun and Armağan Karahanoğlu. 2023. Data sensemaking in self-tracking: Towards a new generation of self-tracking tools. *International Journal of Human–Computer Interaction* 39, 12 (2023), 2339–2360. Publisher: Taylor & Francis.

[25] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic gap in predicting mental wellbeing through passive sensing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*.

[26] Vedant Das Swain and Koustuv Saha. 2024. Teacher, Trainer, Counsel, Spy: How Generative AI can Bridge or Widen the Gaps in Worker-Centric Digital Phenotyping of Wellbeing. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3663384.3663401

[27] Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *Comput. Surveys* 56, 4 (2023), 1–41. Publisher: ACM New York, NY, USA.

[28] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

[29] Brenda Dervin. 1992. In *From the mind's eye of the "user": The sense-making qualitative-quantitative methodology*. Qualitative research in information management, Vol. 1. 24.

[30] Brenda Dervin and others. 1983. An overview of sense-making research: Concepts, methods, and results to date. (1983). Publisher: Author.

[31] Carlo C DiClemente, Angela S Marinilli, Manu Singh, and Lori E Bellino. 2001. The role of feedback in the process of health behavior change. *American journal of health behavior* 25, 3 (2001), 217–227. Publisher: PNG Publications and Scientific Research Limited.

[32] Daniel Durstewitz, Georgia Koppe, and Andreas Meyer-Lindenberg. 2019. Deep neural networks in psychiatry. *Molecular psychiatry* 24, 11 (2019), 1583–1598. Publisher: Nature Publishing Group UK London.

[33] J Engels. 2003. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology* 56, 10 (Oct. 2003), 968–976. https://doi.org/10.1016/S0895-4356(03)00170-7

[34] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4 (Dec. 2020), 126:1–126:38. https://doi.org/10.1145/3432231

[35] 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).

[36] Ivan R Felix, Luis A Castro, Luis-Felipe Rodriguez, and Oresti Banos. 2019. Mobile sensing for behavioral research: A component-based approach for rapid deployment of sensing campaigns. *International Journal of Distributed Sensor Networks* 15, 9 (2019), 1550147719874186. Publisher: SAGE Publications Sage UK: London, England.

[37] Danna N Greenberg. 1995. Blue versus gray: A metaphor constraining sensemaking around a restructuring. *Group & Organization Management* 20, 2 (1995), 183–209. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[38] Rebecca Gulotta, Jodi Forlizzi, Rayoung Yang, and Mark Wah Newman. 2016. Fostering engagement with personal informatics systems. In *Proceedings of the 2016 ACM conference on designing interactive systems*. 286–300.

[39] Andreas Habermacher, Argang Ghadiri, and Theo Peters. 2020. Describing the elephant: a foundational model of human needs, motivation, behaviour, and wellbeing. https://doi.org/10.31234/osf.io/dkbqa

[40] Yunjo Han, Panyu Zhang, Minseo Park, and Uichin Lee. 2024. Systematic Evaluation of Personalized Deep Learning Models for Affect Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (Nov. 2024), 1–35. https://doi.org/10.1145/3699724

[41] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.

[42] Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, and others. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984* (2024).

[43] Jason L Huang, Mengqiao Liu, and Nathan A Bowling. 2015. Insufficient effort responding: examining an insidious confound in survey data. *Journal of Applied Psychology* 100, 3 (2015), 828. Publisher: American Psychological Association.

[44] Nicholas C. Jacobson, Damien Lekkas, Raphael Huang, and Natalie Thomas. 2021. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. *Journal of Affective Disorders* 282 (March 2021), 104–111. https://doi.org/10.1016/j.jad.2020.12.086

[45] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Multimodal Autoencoder: A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 202–208. https://doi.org/10.1109/ACII.2017.8273601

[46] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers?. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE Computer Society, 38–43. https://doi.org/10.1109/FMSys62467.2024.00011

[47] Qi Jiang, Guodao Sun, Yue Dong, and Ronghua Liang. 2021. DT2VIS: A Focus+Context Answer Generation System to Facilitate Visual Exploration of Tabular Data. *IEEE Computer Graphics and Applications* 41, 5 (2021), 45–56. https://doi.org/10.1109/MCG.2021.3097326

[48] Armağan Karahanoğlu and Aykut Coşkun. 2024. Designing for data sensemaking practices: a complex challenge. *Interactions* 31, 4 (2024), 28–31. Publisher: ACM New York, NY, USA tex.issue_date: July - August 2024.

[49] Bangul Khan, Hajira Fatima, Ayatullah Qureshi, Sanjay Kumar, Abdul Hanan, Jawad Hussain, and Saad Abdullah. 2023. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials & Devices* 1, 2 (2023), 731–738. Publisher: Springer.

[50] 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. https://doi.org/10.48550/arXiv.2401.06866 arXiv:2401.06866 [cs].

[51] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

[52] Lisa Krizan. 1999. *Intelligence essentials for everyone*. Joint Military Intelligence College. Number: 6.

[53] Jenny Kunz and Marco Kuhlmann. 2024. Properties and Challenges of LLM-Generated Explanations. https://doi.org/10.48550/arXiv.2402.10532 arXiv:2402.10532 [cs].

[54] Amanda Lazar, Christian Koehler, Theresa Jean Tanenbaum, and David H Nguyen. 2015. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 635–646.

[55] Ha Le, Rithika Lakshminarayanan, Jixin Li, Varun Mishra, and Stephen Intille. 2024. Collecting self-reported physical activity and posture data using audio-based ecological momentary assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3 (Sept. 2024). https://doi.org/10.1145/3678584 Number of pages: 35 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 111 tex.issue_date: September 2024.

[56] Ha Le, Veronika Potter, Rithika Lakshminarayanan, Varun Mishra, and Stephen Intille. 2025. Feasibility and Utility of Multimodal Micro Ecological Momentary Assessment on a Smartwatch. *CHI Conference on Human Factors in Computing Systems (CHI '25)* (2025). https://doi.org/10.1145/3706598.3714086

[57] Younghun Lee, Sungchul Kim, Tong Yu, Ryan A. Rossi, and Xiang Chen. 2024. Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models. https://doi.org/10.48550/arXiv.2402.14195 arXiv:2402.14195 [cs].

[58] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, 9459–9474.

[59] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving Quantitative Reasoning Problems with Language Models. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 3843–3857. https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html

[60] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.

[61] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*. 405–414.

[62] Jiachen Li, Justin Steinberg, Xiwen Li, Akshat Choube, Bingsheng Yao, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2024. Vital insight: Assisting experts' sensemaking process of multi-modal personal tracking data using visualization and LLM. *arXiv preprint arXiv:2410.14879* (2024).

[63] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, and others. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).

[64] Peng Liao, Walter Dempsey, Hillol Sarker, Syed Monowar Hossain, Mustafa Al'Absi, Predrag Klasnja, and Susan Murphy. 2018. Just-in-time but not too much: Determining treatment timing in mobile health. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 4 (2018), 1–21. Publisher: ACM New York, NY, USA.

[65] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, and others. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173* (2024).

[66] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (March 2018), 1–21. https://doi.org/10.1145/3191753

[67] Yanick Xavier Lukic, Gisbert Wilhelm Teepe, Elgar Fleisch, and Tobias Kowatsch. 2022. *Breathing as Input Modality in a Gameful Breathing Training App: Development and Evaluation of Breeze 2 (Preprint)*. preprint. JMIR Serious Games. https://doi.org/10.2196/preprints.39186

[68] Alan Lundgard and Arvind Satyanarayan. 2022. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 1073–1083. https://doi.org/10.1109/TVCG.2021.3114770

[69] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128* (2022).

[70] Lena Mamykina, Elizabeth M Heitkemper, Arlene M Smaldone, Rita Kukafka, Heather J Cole-Lewis, Patricia G Davidson, Elizabeth D Mynatt, Andrea Cassells, Jonathan N Tobin, and George Hripcsak. 2017. Personal discovery in diabetes self-management: discovering cause and effect using self-monitoring data. *Journal of biomedical informatics* 76 (2017), 1–8. Publisher: Elsevier.

[71] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, and others. 2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464* (2024).

[72] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. ACM Press, 1708–1716. https://doi.org/10.1145/3267305.3267537

[73] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting receptivity for mHealth interventions in the natural environment. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–24. Publisher: ACM New York, NY, USA.

[74] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Transactions on Computing for Healthcare (HEALTH)* 1, 2 (April 2020), 1–30. https://doi.org/10.1145/3361562 Publisher: Association for Computing Machinery.

[75] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the reproducibility of physiological stress detection models. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 4 (2020), 1–29. Publisher: ACM New York, NY, USA.

[76] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47. Publisher: Annual Reviews.

[77] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. 2024. Momentary stressor logging and reflective visualizations: Implications for stress management with wearables. In *Proceedings of the CHI conference on human factors in computing systems*. 1–19.

[78] Jasmin Niess and Paweł W Woźniak. 2018. Supporting meaningful personal fitness: The tracker goal evolution model. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[79] Donald A Norman and Daniel Gureasko Bobrow. 1975. *On the role of active memory processes in perception and cognition.* Center for Human Information Processing, Department of Psychology ....

[80] OpenAI. [n. d.]. OpenAI documentation: Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering

[81] OpenAI. 2024. GPT-4o System Card. https://doi.org/10.48550/arXiv.2410.21276 arXiv:2410.21276 [cs].

[82] Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. https://doi.org/10.48550/arXiv.2403.19857 arXiv:2403.19857 [cs].

[83] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[84] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, Thilo Deckersbach, Alex "Sandy" Pentland, and others. 2017. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of medical Internet research* 19, 3 (2017), e75. Publisher: JMIR Publications Toronto, Canada.

[85] Aare Puussaar, Adrian K. Clear, and Peter Wright. 2017. Enhancing Personal Informatics Through Social Sensemaking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6936–6942. https://doi.org/10.1145/3025453.3025804

[86] Yan Qu and Derek L Hansen. 2008. Building shared understanding in collaborative sensemaking. In *Proceedings of CHI 2008 sensemaking workshop*.

[87] Shriti Raj, Joyce M. Lee, Ashley Garrity, and Mark W. Newman. 2019. Clinical data in context: Towards sensemaking tools for interpreting personal health data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1 (2019), 1–20. Publisher: ACM New York, NY, USA.

[88] Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. CataractBot: An LLM-powered expert-in-the-loop chatbot for cataract patients. *arXiv preprint arXiv:2402.04620* (2024).

[89] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal tracking as lived informatics. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1163–1172.

[90] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[91] Koustuv Saha, Manikanta D. Reddy, Vedant das Swain, Julie M. Gregg, Ted Grover, Suwen Lin, Gonzalo J. Martinez, Stephen M. Mattingly, Shayan Mirjafari, Raghu Mulukutla, Kari Nies, Pablo Robles-Granda, Anusha Sirigiri, Dong Whi Yoo, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Sidney K. D'Mello, Anind K. Dey, Kaifeng Jiang, Qiang Liu, Gloria Mark, Edward Moskal, Aaron Striegel, and Munmun de Choudhury. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 178–184. https://doi.org/10.1109/ACII.2019.8925479 ISSN: 2156-8111.

[92] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 671–676.

[93] Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. Association for Computing Machinery, 400–408. https://doi.org/10.1145/3242969.3242985

[94] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1 (2008), 1–32. Publisher: Annual Reviews.

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

[96] Konstantin R Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W Woźniak. 2024. Narrating fitness: Leveraging large language models for reflective fitness tracker data interpretation. In *Proceedings of the CHI conference on human factors in computing systems*. 1–16.

[97] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 645–654. https://doi.org/10.1145/3616855.3635752

[98] ML Tlachac, Ricardo Flores, Miranda Reisch, Katie Houskeeper, and Elke A Rundensteiner. 2022. DepreST-CAT: Retrospective smartphone call and text logs collected during the covid-19 pandemic to screen for mental illnesses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–32. Publisher: ACM New York, NY, USA.

[99] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 74952–74965. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

[100] Sudip Vhaduri and Christian Poellabauer. 2017. Design factors of longitudinal smartphone-based health surveys. *Journal of Healthcare Informatics Research* 1 (2017), 52–91. Publisher: Springer.

[101] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[102] Yun Wang, Zhitao Hou, Leixian Shen, Tongshuang Wu, Jiaqi Wang, He Huang, Haidong Zhang, and Dongmei Zhang. 2022. Towards Natural Language-Based Visualization Authoring. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–11. https://doi.org/10.1109/TVCG.2022.3209357

[103] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, and others. 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.

[104] Haoyi Xiong, Jiang Bian, Sijia Yang, Xiaofei Zhang, Linghe Kong, and Daqing Zhang. 2023. Natural Language based Context Modeling and Reasoning for Ubiquitous Computing with Large Language Models: A Tutorial. https://doi.org/10.48550/arXiv.2309.15074 arXiv:2309.15074 [cs].

[105] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–27. https://doi.org/10.1145/3448107

[106] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, and others. 2023. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34. Publisher: ACM New York, NY, USA.

[107] Xuhai Xu, Jennifer Mankoff, and Anind K Dey. 2021. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. *CCF Transactions on Pervasive Computing and Interaction* 3 (2021), 344–366. Publisher: Springer.

[108] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, and others. 2022. GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization. *Advances in Neural Information Processing Systems* 35 (2022), 24655–24692.

[109] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (Nov. 2024), 153:1–153:29. https://doi.org/10.1145/3699765

[110] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems* 2, 1 (Jan. 2025), 1–5. https://doi.org/10.1038/s44401-024-00004-1 Publisher: Nature Publishing Group.

[111] 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.
[112] Pengyi Zhang and Dagobert Soergel. 2014. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology* 65, 9 (2014), 1733–1756. Publisher: Wiley Online Library.
[113] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* 5, 5 (2021), 726–742. Publisher: IEEE.