

# Feasibility of using a multi-agent LLM system to correct annotations and support low-effort activity labeling

HA LE, Northeastern University, USA

AKSHAT CHOUBE, Northeastern University, USA

VARUN MISHRA, Northeastern University, USA

STEPHEN INTILLE, Northeastern University, USA

Accurate measurement of behaviors is critical for research in human-computer interaction, ubiquitous computing, and personal health informatics, because it underpins many health tracking and intervention systems. Most data collection studies, however, still rely on participants' self-reports and manual annotations, which often under- or over-estimate activity duration and type, and require substantial effort from researchers to clean and validate. An automated system that can combine passive sensing data with participants' self-reports might detect inconsistencies and suggest corrections. We introduce GLOSS4HAR, a multi-agent LLM-based system designed to mimic human sensemaking and assist researchers in cleaning and refining activity annotations. We demonstrate the potential of GLOSS4HAR in two key tasks: (1) correcting and reconciling participant self-annotations, and (2) corroborating passive sensing data with different forms of lightweight self-reports to generate accurate activity timelines. Our evaluation shows that GLOSS4HAR improves annotation quality by up to 9.9% in F1 score and can reconstruct activity timelines that align with human annotations at 75–92% F1. Based on our findings, we discuss the implications of our work on the next generation of activity annotation systems through human-AI collaboration.

## ACM Reference Format:

Ha Le, Akshat Choube, Varun Mishra, and Stephen Intille. 2026. Feasibility of using a multi-agent LLM system to correct annotations and support low-effort activity labeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1 (April 2026), 36 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Accurately measuring daily activities and postures is critical for applications in health, wellness, sports, and behavioral research [10, 60, 65, 82, 83]. Many studies rely on participants' self-reports, which are often affected by recall bias, missingness, and reporting burden [55, 57, 80, 89]. Passive sensing and human activity recognition (HAR) models provide an alternative by continuously inferring activities [8, 50, 87], enabling applications such as real-time interventions and behavior monitoring [3, 56, 64, 109]. However, these systems still depend on high-quality labeled data, which remains difficult to collect in real-world settings.

Collecting accurate activity labels in the wild is challenging. Egocentric video can provide detailed ground truth but raises serious privacy concerns and requires substantial manual effort, limiting scalability [13, 31, 44]. As a result, researchers often rely on self-reports, including retrospective recall [34, 48, 49, 68], sensor-assisted recall [6, 74, 79], in-the-moment reporting [47, 54, 55], and simplified summaries such as activity lists or free-form

---

Authors' addresses: Ha Le, le.ha1@northeastern.edu, Northeastern University, Boston, USA; Akshat Choube, choube.a@northeastern.edu, Northeastern University, Boston, USA; Varun Mishra, v.mishra@northeastern.edu, Northeastern University, Boston, USA; Stephen Intille, s.intille@northeastern.edu, Northeastern University, Boston, USA.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2474-9567/2026/4-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

narration [35, 66, 80]. While these approaches vary in effort and expressiveness, they introduce inconsistencies in temporal precision and completeness, making it difficult to generate consistent and temporally aligned activity annotations.

Depending on study design, participant demographics, and available resources, researchers adopt a range of annotation strategies, from high-effort, detailed methods to low-effort, lightweight ones. A common approach is recall-based self-reporting, where participants reflect on and record their daily activities, often with contextual cues such as sensor data [36]. This typically requires reconciling memory with data to produce fine-grained, minute-level annotations. To reduce burden, some studies use in-the-moment ( $\mu$ EMA) reporting, prompting participants throughout the day to record their current activity [47, 54, 55]. While faster, this can result in missing data due to contextual factors such as social interactions or physically demanding tasks [55, 70, 76]. Other approaches further reduce effort by asking participants to list activities chronologically, with or without approximate durations [80], or to provide open-ended text or voice summaries [35, 66]. Although these methods improve usability and engagement, they vary in temporal precision, completeness, and structure. Researchers often combine multiple approaches to improve compliance, but these sources do not always align or provide equivalent information, complicating downstream processing. **This diversity motivates approaches that can integrate and reconcile multiple low-effort self-reports with passive sensing data to produce consistent, temporally aligned activity timelines.**

These challenges are compounded by the need for large, high-quality datasets to train HAR models. Many existing datasets are collected in controlled laboratory settings and do not generalize well to real-world environments [15, 21, 103]. Even small labeling errors can significantly degrade model performance [51], requiring researchers to manually clean and realign annotations with sensor data—a process that is time-consuming and inconsistent [37, 46].

These challenges highlight the need for **automated systems that can identify and correct annotation errors**. Even small labeling errors can significantly degrade model performance [51], and researchers often spend considerable time manually reviewing and refining annotations to align them with sensor data [37, 46]. This process is labor-intensive, difficult to scale, and prone to inconsistencies across annotators. Automating annotation correction could therefore improve data quality while reducing the burden on researchers.

At the same time, there is a need for **low-burden data collection methods** that reduce reliance on detailed retrospective annotations. High-effort annotation protocols limit scalability and long-term participation [55], motivating the use of lightweight self-report methods such as  $\mu$ EMA, activity lists, or free-form summaries [35, 66, 80]. These low-effort self-report, however, are often incomplete, noisy, and temporally imprecise when used in isolation. This creates an opportunity for systems that can **combine multiple low-effort self-reports with passive sensing data** to reconstruct detailed and temporally aligned activity timelines, while maintaining low participant burden.

In this work, we introduce **GLOSS4HAR**, a multi-agent LLM system that supports human activity annotation by triangulating passive sensing data with multiple forms of low-effort self-reports. We evaluate the system on a one-week, free-living dataset collected using ACAI [6], which includes passive sensing streams and participants' retrospective annotations. We study two complementary tasks: (1) correcting errors in existing annotations and (2) reconstructing detailed activity timelines from low-effort self-reports. Using participants' detailed annotations, we derive multiple forms of simplified self-reports (e.g.,  $\mu$ EMA, activity lists with or without temporal information) and evaluate how well the system can combine these inputs with sensor data to recover timelines comparable to researcher-refined ground truth.

We address the following research questions:

- **RQ1:** Can LLMs identify and correct incorrect retrospective activity annotations made by the participants in a research study, reducing researcher burden?

- **RQ2:** Can LLMs reconstruct detailed timelines from low-effort self-reports combined with passive sensing data?

Our contributions are:

- We introduce GLOSS4HAR, a multi-agent LLM system for activity annotation, and demonstrate its ability to flag and correct errors in retrospective reports.
- We evaluate its ability to support low-effort data collection by combining passive sensing data with common self-report formats, including  $\mu$ EMA and activity lists with or without temporal information.
- We outline how such systems could support real-time, human-in-the-loop annotation by generating context-aware  $\mu$ EMA prompts and incorporate open-ended end-of-day narration to improve data quality with minimal burden.

**Paper structure.** Section 2 reviews related work. Section 3 describes the ACAI dataset and our revision of participant annotations. Section 4 presents the original GLOSS system and our modifications for HAR (GLOSS4HAR). Section 5 introduces the annotation correction and timeline reconstruction tasks, along with baselines and evaluation metrics. Section 6 reports results. Section 7 discusses potential real-time, low-burden use cases. Section 8 provides discussion, Section 9 outlines limitations, and Section 10 concludes.

## 2 BACKGROUND

Our work builds upon the prior literature on activity data collection methods, the increased use, potential and limitations of human-AI collaborative annotation systems, and the use of LLMs in sensing research.

### 2.1 Collecting annotations for human activity recognition datasets

Human activity recognition (HAR) is the task of identifying activities and postures using data from wearables, sensors, or cameras [1, 8, 33]. Accurate HAR models might benefit clinicians, self-trackers, and designers of just-in-time intervention systems. However, the quality of HAR models is highly dependent on the quality of their training datasets [51]. Most HAR datasets are collected in controlled laboratory settings [21], where participants perform fixed activity sequences while researchers log timestamps manually. Models trained on such data often fail to generalize to free-living settings, where activities are far more diverse and fast-changing [2, 103].

Collecting annotations from free-living people is challenging [13]. Ego-centric cameras mounted on the head or chest [45, 84] can yield accurate labels but are tedious to annotate, costly, and raise serious privacy concerns for participants and bystanders [44]. A common alternative is participant self-report, either *in-the-moment* via ecological momentary assessment (EMA or  $\mu$ EMA) [39, 54, 55, 85] or *after-the-fact* recall at the end of the day [35, 74, 79].

EMA can be intrusive and lacks temporal granularity [76], while recall is vulnerable to memory bias and is cognitively demanding, especially when activities overlap [6, 14, 51, 79]. When errors happen, researchers must reconcile self-reports with sensor data by hand – a process that is time-consuming and difficult to recover from once participants have left the study [18, 41]. This highlights the need for automated systems that can detect inconsistencies, flag errors, and suggest corrections in real time [35, 62].

### 2.2 Human-AI collaborative annotation systems

As discussed in the previous sections, both participants and automated systems face limitations: humans bring contextual understanding but often struggle with recall and consistency, and AI models can process sensing data at scale but lack situational awareness. Combining these complementary strengths holds promise, but effective collaboration requires careful design. Studies show that human-AI teams can sometimes perform worse than either alone when trust is miscalibrated—either through under-reliance (ignoring useful suggestions) or over-reliance (accepting incorrect ones) [7, 12, 42, 52, 91].

For the annotation task, participants can refine or correct AI-generated predictions using their contextual knowledge while AI support simultaneously lowers cognitive effort and speeds up the process. Prior work has demonstrated the value of this approach across domains, including document [26], video [23], image [86], audio [5], and biomedical data [9]. More recently, annotation systems for passive sensing data have included AI-generated labels. For instance, Stojchevska et al. allowed participants to review and edit predicted activity labels at the end of the day [88], and ExtraSensory [92], ACAI [6], and Neupane et al. [74] enabled users to refine model predictions, improving labeling accuracy and reducing effort.

### 2.3 Large Language Models in Passive Sensing Research

Researchers in passive sensing are increasingly leveraging LLMs for their strong commonsense reasoning and world knowledge, applying them to behavioral and health outcomes including depression [17], stress [71], sleep [53, 81], and human activities [32]. Common approaches include converting sensing data into textual representations [61, 90, 104], retrieval-augmented generation over sensor databases [59, 106], and fine-tuning LLMs on raw sensor streams [104, 105, 110]. Researchers have also built LLM-based sensemaking systems to help diverse stakeholders interpret passive sensing data, including VitalInsight [61], Dr. House [105], PhysioLLM [27], PHIA [69], SensorChat [107], and GLOSS [19]. More recently, LLMs have been used to assist participants directly – generating journaling prompts [72, 73], narrative activity summaries [102], smartwatch voice assistance for procedural tracking [3, 4], and voice-based timeline annotation [35].

In the HAR space specifically, LLMs have been applied to improving recognition accuracy [30, 58, 63, 110], explaining model predictions [16, 97, 107], and generating synthetic training data [58]. All of these works, however, assume the correctness of their underlying annotations. Activity annotation is cognitively demanding and burdensome for participants [6, 20, 55], yet the potential of LLMs to reduce this burden remains underexplored. In this work, we adapt GLOSS [19] into GLOSS4HAR to address this gap, demonstrating the feasibility of LLMs for (1) correcting human annotations to improve label quality, and (2) supporting low-effort annotation interfaces such as  $\mu$ EMA, activity lists, and narrations, including through a real-time simulation study.

## 3 ACAI DATASET: OVERVIEW AND ANNOTATION QUALITY

In this section, we provide details on our previous study that generated the dataset used in this work<sup>1</sup>. We describe how we carefully used manual triangulation of passive sensing data self-reports to obtain the ground truth activity labels for free-living people used in this work.

### 3.1 ACAI platform and dataset

The ACAI study used a mobile and smartwatch app for research data collection, capturing both passive sensing data and participant-provided activity annotations (Figure 1) [6]. Every 15 minutes, participants answered  $\mu$ EMA prompts on the smartwatch via speech. The app transcribed speech on-device using a fine-tuned Google Cloud Speech-to-Text model [28] and saved only the transcriptions. The app displayed visualizations of the collected sensing data and  $\mu$ EMA responses to help participants recall and label their activities accurately. The interface allowed participants to review their data at the end of the day and annotate start and end times for activities in the timeline at a five-minute granularity. Although this coarse level of annotation reduced participant burden compared to minute-by-minute labeling, it introduced imprecision in temporal labeling.

The ACAI dataset has two components: data from a one-day study and data from a seven-day free-living study. In the one-day usability dataset, 11 participants wore a Pixel 2 smartwatch that recorded passive sensor data and  $\mu$ EMA in-the-moment activity self-reports [55]. After one day of data collection, participants engaged

<sup>1</sup>We have included the anonymized version of the paper about the ACAI study in the supplemental materials. This paper was recently accepted to IMWUT.

Table 1. Data streams and their sampling frequencies in the ACAI dataset.

Device	Data Stream (Sampling Frequency)
Phone	GPS location (1m), app use (1m), lock/unlock events (1m)
Smartwatch	Step counts (1m), ambient noise classification [29] (5m), heart rate (1m), $\mu$ EMA responses (15m), skin temperature (10s), wrist movement data (10s).



Fig. 1. Mobile interface of ACAI [6]. The top two-thirds of our annotation interface presented participants with their self-reports and passively collected contextual cues to aid with the recall process. The bottom third displayed the annotation timeline and labeling controls.

in a one-hour session in the lab using the ACAI app to annotate their activities from the previous day. In the second portion of the study, 14 participants wore the Pixel 2 smartwatch to collect passive sensing data and  $\mu$ EMA responses once every 15 minutes for seven days, completing annotations in their naturalistic settings. We used the data from both portions of the study for our experiments in this paper.

The ACAI dataset contains passive sensing data from both the smartphone and the smartwatch to complement these active self-reports (see Table 1).

To clarify the relationship between the original ACAI study and this work, participants in the ACAI dataset did not review or correct their annotations. In ACAI, participants collected passive sensing data throughout the day, responded to micro-EMA prompts (e.g., “what are you doing now?”), and completed end-of-day retrospective annotations that included start/end times, posture, and activities. Any correction of these annotations was not part of the original ACAI pipeline. In this work, we first introduce a researcher annotation review step, where our team reviews and refines participants’ retrospective annotations to address errors arising from interface

limitations and recall (Section 3.2). This serves as the reference for RQ1, where we evaluate whether GLOSS4HAR can automate this correction process (Section 5.1). In addition, we simulated low-burden self-report formats, which is entirely new in this work and was not part of the ACAI dataset (Section 5.2). This enables RQ2, where we investigate whether LLMs can reconstruct detailed timelines from low-effort self-reports combined with passive sensing data.

### 3.2 Reviewing activity annotations from the ACAI dataset

In this paper, we use the ACAI dataset because we need a dataset of temporally dense participant annotations where we can also obtain a good estimate of ground truth activity. Having participants wear cameras continuously—and manually labeling the resulting video—was not practical; instead, we use the multimodal sensing data in the ACAI to obtain a good estimate of ground truth.

Errors in activity annotations in the ACAI dataset resulted from recall biases, participants having difficulty in reviewing passive sensing data, or missing details (for instance, exact start and end times of annotations). The original ACAI paper highlights many of these errors presented in the participants’ annotations [6]. To obtain an estimate of ground truth, we manually reviewed and triangulated the following sources of data:

- **Participants’ retrospective annotations.** Participants in the ACAI study reviewed their passive sensing data using the mobile interface and annotated activities they did throughout the day. They had the option to indicate their uncertainty about the annotation.
- **Participants’ in-the-moment activity annotations.** Participants were prompted to answer an open-ended, multimodal  $\mu$ EMA questions about their in-situ activity and postures once every 15 minutes.
- **Smartwatch/smartphone sensing data.** Sensing data from the participants’ smartphones and the provided Pixel Watch 2 smartwatch were collected: location data once every minute, step count data once every minute, ambient noise classification once every minute, skin temperature data once every second, heart rate once every second, phone usage data, and wrist acceleration data at 50 Hz.
- **Thigh sensor.** Participants wore an Actigraph GT9X-Link sensor on their thigh for the entire study duration recording the tri-axial accelerometer data (sampled at 80 Hz). These data were used to predict the participants’ postures (lying down, sitting, standing, walking, and running) once every minute.

We reviewed the annotations in the ACAI dataset and validated the annotations by finding corroborating evidence in the passive sensing data recorded by the smartwatch, smartphone, and thigh sensors. For example, if a participant annotated “walking” but there were no steps recorded and the participant was wearing the watch during this time, the annotation was marked as incorrect. We also looked for temporal misalignment in the annotation, e.g., if the participant annotated “riding bus” from 10:30 am to 11:00 am, but the GPS data indicated that the participant was only traveling from 10:30 am to 10:55 am, the period between 10:55 am to 11:00 am was marked as incorrect.

We only made changes to the annotation if there was **clear evidence** in the data supporting the errors. We identified and fixed three main classes of errors in the annotations: *bunching*, *omission*, and *boundary misalignment*.

- **Bunching.** Bunching occurs when participants merge multiple distinct activity segments into a single annotation. For example, “walking, using phone” (9:00–9:05 am) and “standing” (9:05–9:10 am) may be combined into “walking, standing, using phone” for the full 9:00–9:10 am window. We detect these errors by identifying infeasible label combinations (e.g., “sitting” and “walking” simultaneously) and flag them for correction. Fixes typically involve splitting the annotation into smaller segments (e.g., separating “sitting” and “walking”) using evidence from thigh sensor and step count data.
- **Omission.** Omission occurs when parts of annotations are missing. This includes: (1) short missing segments, often due to oversight or the inability to annotate intervals shorter than five minutes, and (2) missing posture or activity labels. We detect short gaps by identifying change points in thigh sensor

and passive sensing data and insert missing segments when evidence supports a transition. For missing labels, we check whether both posture and activity are present: posture is added from thigh sensor data, and activity is inferred from available signals (e.g.,  $\mu$ EMA responses). When evidence is insufficient, the annotation is labeled “*unknown*.”

- **Boundary misalignment.** Boundary misalignment occurs when start or end times are over- (*overflow*) or under-estimated (*underfill*) by more than one minute, often due to recall bias or five-minute annotation granularity. Overestimation distorts activity durations, while underestimation obscures transitions. Prior work shows that even 15% misalignment can significantly degrade HAR performance [51]. We detect these errors by comparing annotations with passive sensing and thigh sensor data. For example, if “walking” is labeled from 9:00–9:10 am but sensor data indicates 9:02–9:08 am, we adjust boundaries accordingly, extending or trimming segments when evidence supports it.

Despite our best effort to reconcile the participants’ annotations and the underlying data, there are cases where the phone/watch sensing data and the thigh sensor did not provide enough evidence about the correct activities that the participants might have been doing. In these instances, we marked the activity as “*unknown*” and omitted these annotations from subsequent experiments in this paper. Overall, we modified 22.6% of the minutes in the dataset. Among all annotation bouts provided by participants, 5.4% contained bunching errors, 27.6% contained temporal errors, and 33% contained omission errors (primarily due to missing activity labels). 8.4% of the labels contained an “*unknown*” activity; *unknown* annotations were removed from the analysis.

## 4 GLOSS4HAR: SYSTEM OVERVIEW

In this section, we discuss the multi-agent LLM network we chose for our work: GLOSS [19]. We also describe how we adapted the GLOSS implementation to address the problem of correcting activity annotations and triangulating passive sensing data with low-effort self-reports to reconstruct the activity timelines.

### 4.1 GLOSS - Networks of LLMs for multimodal sensing data

GLOSS is an open-source system consisting of multiple LLM agents designed to mimic the process of sensemaking in humans. The sensemaking process in GLOSS involves two cyclic processes: the *information seeking* phase focusing on retrieving information from the datasets and processing the raw data into more a understandable format; and the *sensemaking* phase focusing on triangulating and interpreting the results from multiple data streams, as well as presenting the final results. GLOSS’s network includes eight LLM agents (Figure 2): in the *information seeking* phase, an **action plan generation agent** and **next step agent** coordinate query planning, and an **information seeking agent** and **database manager agent** retrieve relevant data, which a **code generation agent** then processes by writing and executing Python scripts using pre-defined helper functions; and in the *sensemaking* phase, **local** and **global sensemaking agents** iteratively interpret the retrieved data into a growing understanding of the query, until the system determines the understanding is sufficient and hands off to a **presentation agent** to format the final response. GLOSS has outperformed a RAG-based system in responding to open-ended questions about passive sensing data, both in terms of accuracy and consistency [19].

### 4.2 Extending GLOSS for the problem of activity annotation

We made changes to GLOSS to adapt the method to the task of correcting activity annotations; it does this correction by triangulating passive sensing data with multiple forms of low-burden self-report to generate a comprehensive activity timeline. We introduced three key modifications to GLOSS: (1) refining the user prompt to better handle activity annotations; (2) adding an adaptive database that stores past corrections generated by GLOSS4HAR (i.e., earlier-in-time corrections made by the system itself), enabling the model to capture recurring

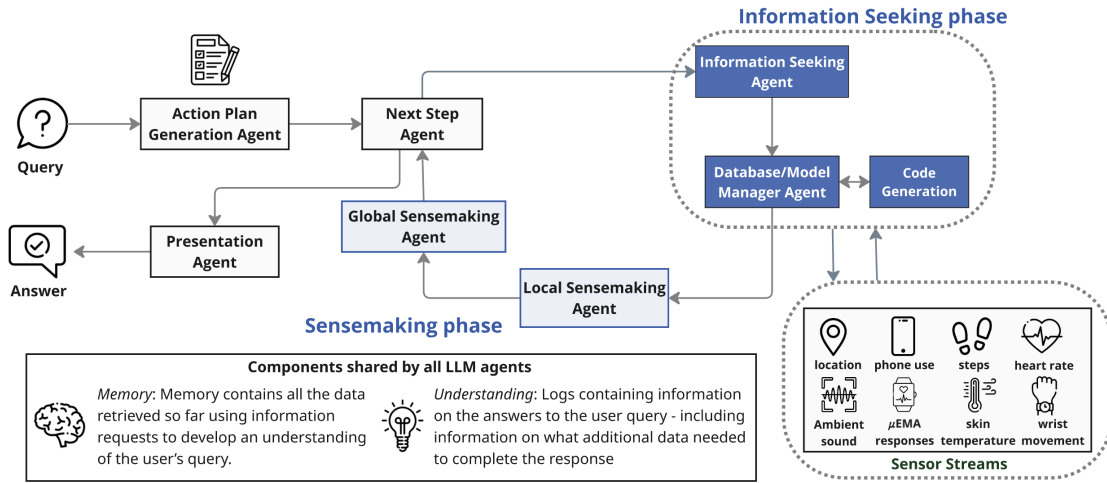


Fig. 2. Networks of LLM agents and their interactions in GLOSS4HAR, and the list of sensor streams available in the ACAI dataset (image is the adaptation of GLOSS for the task of activity annotation [19]).

patterns between posture and contextual activities; and (3) ensuring that GLOSS’s final output follows the activity mapping and adheres to a fixed format.

**4.2.1 Adapting system prompt for activity annotations.** We show the system prompt we used for the annotation correction task and the activity timeline generation task in the supplementary materials of this paper.

In the annotation correction task, we provided the LLM with the participants’ self-reported activity annotation (one annotation at a time), the timestamp of the annotations, and the definitions of the three types of errors (Section 3.2) to look for. For this task, we allowed GLOSS4HAR to access all the passive sensing data available in the ACAI dataset, *except for accelerometer data for the thigh sensor* (because we use the thigh sensor as ground truth in our experiment). We included the full list of sensor data and description for database management agent in Appendix D.

For the activity timeline generation task, we used the same implementation of GLOSS4HAR as for the correction task, except for the system prompt. We prompted GLOSS4HAR to generate an activity timeline in one-hour bouts—for example, asking it to generate a list of postures and activities a participant engaged in from 8 am to 9 am. For the  $\mu$ EMA condition, participants self-report are not included in the system prompt, but as a separate database for GLOSS4HAR to query. For the other conditions, the participants’ self-reports were included in the system prompt.

**4.2.2 Retaining memory of past annotations for better context.** Although passive sensing data provide rich information about a person’s contexts, many activities can look similar from sensor signals alone (e.g., sitting while reading vs. sitting while eating). As a result, participant input is still necessary to ensure accurate predictions. Past self-reports are especially valuable because they give the LLM additional context about a participant’s behaviors and habits. Thus, we added an adaptive database into GLOSS: the system keeps track of results of previous runs, stores them in a database, and retrieves the database each time it makes a correction. For example, if the system repeatedly sees that a participant’s “standing” label around 7:30 am is paired with “making breakfast,” GLOSS4HAR can learn from this correction and use it to improve future predictions at similar times.

Table 2. List of postures and activities passed to the *Presentation agent* for generating predictions.

Postures	Activities
sitting, standing, lying down, reclining, upright	video gaming, walking, stair climbing, getting ready, driving, bicycling, vigorous bicycling, aerobics, cleaning, cooking, laundry, playing with pet, listening to music, watching movies/TV, studying, reading, riding in car, riding train, riding bus, playing musical instruments, attending meeting, computer using, phone using, running, getting dressed, grooming, using bathroom, eating, talking, strength training, washing dishes, carrying groceries, putting away groceries, shopping, making bed, packing/unpacking, sleeping, playing sports

**4.2.3 Ensuring final output follows activity mapping and fixed format.** GLOSS was originally designed as an open-ended query interface. In the context of human activity recognition and behavioral sciences, however, researchers are interested in a defined and structured set of labels. Thus, we constrained GLOSS4HAR to generate labels from a predefined set of relevant activities, rather than producing arbitrary ones. To ensure consistency and robustness, we provided the list of predefined posture and activity labels (Table 2) —along with the desired output format, to the *Presentation agent*. Our design aims to support the integration of our system to more structured activity recognition or intervention systems that require fixed format. A structured output also makes the validation process easier. The list of activity labels, however, can be extended or made open-ended based on the requirements of the study.

### 4.3 Implementation

We extended the open-source implementation of GLOSS for our system [19], using the LangChain and AutoGen frameworks to manage interactions between LLM agents.

For our experiments, we used GLOSS with two OpenAI large language models: gpt-4o and the gpt-oss:20b model. At the time of the study, GPT-4o was the most recent model released by OpenAI and was available through their cloud service and API. Its performance has been validated across a wide range of benchmarks, including reasoning and coding tasks, and it was also the model used in the original GLOSS implementation [19].

To complement this, we also selected a smaller, open-source model that can run locally on a GPU. The gpt-oss family includes models with 20 billion and 120 billion parameters; we chose the 20b version. Using this small model is valuable because it can be run locally, and thus in a privacy-preserving way; it might also be amenable to future real-time mobile applications. In our GPT-4o API calls, we set *temperature* = 0 and *top<sub>p</sub>* = 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. The gpt-oss model was run locally on a machine with an Nvidia RTX 5090 GPU. For the gpt-oss model, we set the reasoning level to be low for the coding agent, and medium for the other agents.

In our experiments, we stored the raw passive sensing data in CSV format and connected these to the *Database Management Agent* within GLOSS4HAR. We will release our implementation of GLOSS4HAR and code for the experiments to enable reproducibility.

## 5 ANALYSIS PLAN

We present our analysis plan for the annotation correction and activity timeline generation tasks, including baseline and model choice, and the metrics. We also detail our process of simulating low-effort activity annotations by using the participants' annotations in the ACAI dataset.

### 5.1 Baseline, analysis plan and ground truths for annotation correction task (RQ1)

To evaluate the activity annotation correction task (RQ1), we implemented a baseline model using a Retrieval-Augmented Generation (RAG) approach that is commonly applied in prior passive sensing studies [59, 106]. Following this prior research, we kept raw sensor data in natural language format and stored them in CSV files. The RAG model then retrieved relevant text segments based on each query to generate annotation corrections. We showed an example of how the RAG model works in Appendix H.

We evaluated annotation correction performance by comparing three sets of annotations—participants’ original annotations, RAG-corrected annotations, and GLOSS4HAR-corrected annotations—against researcher-corrected annotations, representing the human-inferred ground truth. Our goal is to assess how closely LLM-based systems perform to researcher-level corrections. In addition to alignment accuracy metrics, we also report the proportion of data segments successfully corrected by GLOSS4HAR and RAG to assess correction efficacy.

### 5.2 Simulation of self-reports and analysis plan for activity timeline generation task (RQ2)

For the activity timeline generation task (RQ2), we evaluated how GLOSS4HAR performed under four conditions: three different forms of participant self-report and one baseline condition without any self-report input. We evaluated the alignment between the GLOSS-generated annotations and three reference points: locomotion data detected by thigh-mounted sensors, participants’ original annotations, and researcher-corrected annotations.

*5.2.1 Simulation of low-effort activity annotations.* Depending on the study design and target population, researchers adopt different strategies for collecting activity annotations. Recall-based self-reporting requires participants to reflect on and record their daily activities with the aid of contextual cues, reconciling memory with sensor data to produce detailed annotations — a process that is both time-consuming and cognitively demanding [25, 36]. In-situ reporting reduces memory reliance by prompting participants throughout the day [39, 47, 54, 55], but suffers from missing data when participants are unable or unwilling to respond [55, 70, 76]. Simpler retrospective approaches ask participants to list activities chronologically with or without approximate durations [80], reducing burden at the cost of temporal precision. In this work, we explore how LLMs can support this range of annotation methods by triangulating self-reports with passive sensing data to generate detailed activity annotations, enabling researchers to diversify data collection while maintaining a cohesive annotation schema and potentially improving engagement and retention [94]. To evaluate this potential, we simulated four commonly used annotation approaches:

- **No self-reporting:** Only passive sensing data is collected, eliminating interruption and recall burden but making ground-truth labels difficult to obtain without an objective observation method such as egocentric cameras [24, 95, 96].
- **EMA/ $\mu$ EMA (every 15 minutes):** Smartphone or smartwatch surveys capture accurate, in-the-moment labels with high ecological validity, and participants can sustain high compliance for up to 6 months to a year using micro-interaction  $\mu$ EMA [75]. However, these methods suffer from contextual bias [76] — participants may not respond in certain environments such as classrooms or meetings. We used  $\mu$ EMA data collected from the ACAI study once every 15 minutes.
- **List of activities (with approximate time):** Participants provide a sequence of activities with coarse temporal information (e.g., approximate durations) rather than precise boundaries [35, 43, 49, 88]. We simulate this condition by transforming the original annotations through the following steps: (1) *Information loss:* remove posture labels and simplify activity names, and convert precise start/end timestamps into durations; (2) *Temporal jitter:* perturb each activity duration by 0–40% [51]; (3) *Order perturbation:* randomly swap up to 20% of adjacent activity pairs. This results in a temporally ordered list of activities with approximate durations that are imprecise.



Fig. 3. Simulation of low-effort self-reports. We introduce noise and remove contextual details (removing postures, approximating start/end times with durations, swapping activity order, and jittering durations) to better reflect real-world self-reporting behavior and evaluate system robustness under imperfect or unreliable participant input.

- **List of activities (no time):** Participants provide only an ordered list of activities without any temporal information. Starting from the approximate-time condition, we further apply: (1) *Remove temporal information*: discard all durations. The final output is a sequence of activities with no explicit timing, preserving ordering but inheriting temporal inaccuracies from earlier steps. We showed an example of how the simulation process works in Figure 3.

### 5.3 Metrics calculation and justifications

We evaluated both tasks using five metrics for multi-label time series prediction: exact match ratio, F1 score, Jaccard index, fixed amount (minutes), and misalignment analysis. Exact match (accuracy) is the strictest measure, requiring all labels to match per minute—capturing precise correctness but ignoring partial agreement. F1 balances precision and recall, accounting for both false positives and false negatives. The Jaccard index measures overlap between predicted and true label sets, rewarding partial agreement while penalizing mismatches. Fixed amount quantifies the average minutes per day that GLOSS4HAR correctly fixes or incorrectly modifies. Misalignment analysis provides a segment-level view of temporal and structural errors, including insertions, deletions, boundary misalignments, and fragmentation or merging [100]. Table 3 summarizes these metrics. Together, they capture complementary aspects of performance—from exact correctness to partial agreement—and reveal different types of labeling and temporal errors.

Table 3. Evaluation metrics for assessing annotation quality and temporal alignment.

Metric	What It Measures	Why It Is Useful
<b>Exact Match Ratio (Subset Accuracy)</b>	Minutes where predicted and true labels are <i>identical</i> .	Strict correctness at the minute level (exact set match).
<b>F1 Score</b>	Harmonic mean of precision and recall.	Balances false positives/negatives; robust to imbalance.
<b>Jaccard Index</b>	Overlap between predicted and true label sets.	Captures <i>partial agreement</i> in multilabel settings.
<b>Fixed Amount</b>	Minutes correctly fixed, unchanged, or incorrectly modified.	Measures error detection and correction performance.
<b>Misalignment Analysis [100]</b>	Segment differences: <ul style="list-style-type: none"> <li>• <i>Insertion</i>: extra label</li> <li>• <i>Deletion</i>: missed label</li> <li>• <i>Under/Overfill</i>: boundary errors</li> <li>• <i>Fragment/Merge</i>: split or merged segments</li> </ul>	Identifies temporal and structural errors.

## 6 EVALUATION RESULTS

Our results from both tasks show GLOSS4HAR’s potential to identify and fix annotating errors and to triangulate passive sensing data with minimal participants’ self-report to generate a more accurate activity timeline.

### 6.1 GLOSS4HAR efficacy in flagging and fixing incorrect activity annotations (RQ1)

We present both quantitative and qualitative results from the correction task, demonstrating the ability of GLOSS4HAR to significantly outperform RAG and improve annotation quality.

Table 4. Comparison of annotation correction performance. Note that “Correct” and “Incorrect” fixes are not mutually exclusive—for example, a single fix can simultaneously add a correct activity label (Correct) and remove a correct label (Incorrect).

	Model	Exact Match	F1 Score	Jaccard Index	Average fixed amount (min/day)		
					Correct	Unchanged	Incorrect
Original Annotations	—	67.8	85.9	80.0	—	809.9	—
RAG corrections	gpt-oss 20b	70.4	88.7	84.8	33.7	772.5	3.9
	gpt-4o	74.2	91.7	87.8	58.1	747.4	4.4
GLOSS corrections	gpt-oss 20b	75.7	93.4	88.6	108.7	693.3	15.2
	gpt-4o	77.1	93.6	89.0	125.9	653.5	27.3

**6.1.1 Quantitative findings.** Results from the correction tasks demonstrate GLOSS4HAR’s effectiveness in flagging and correcting potentially incorrect annotations (Table 4). Compared to the RAG-based models, GLOSS4HAR achieved substantially higher performance across all metrics. Applying RAG corrections led to modest improvements: gpt-oss increased exact match to 70.4% and F1 to 88.7%, while gpt-4o reached 74.2% exact match and 91.7% F1. The corresponding average fixed amounts show that most minutes remained unchanged, with only a small number of corrections labeled as correct (33.7–58.1 min/day) or incorrect (3.9–4.4 min/day). In contrast, GLOSS corrections achieved the highest improvements, with gpt-oss yielding 75.7% exact match and 93.4% F1, and gpt-4o

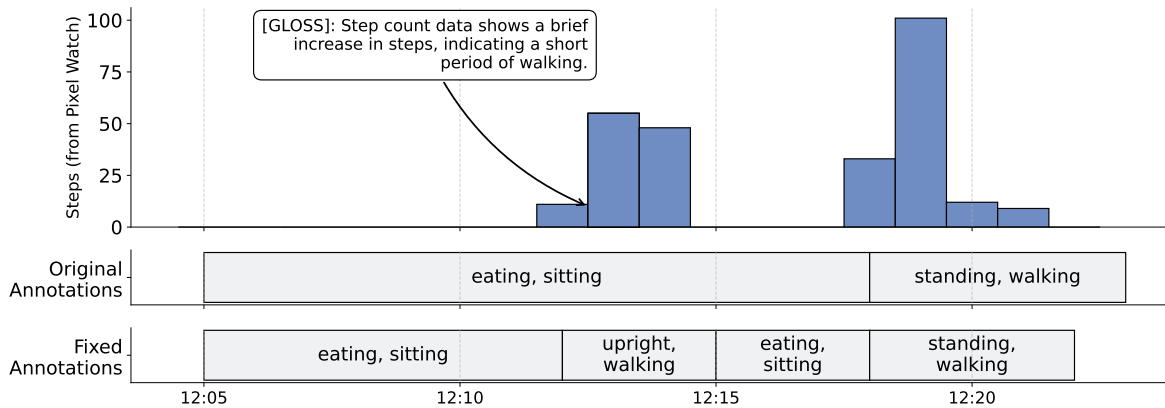


Fig. 4. An example of how GLOSS4HAR can spot a *merge* error (participants **missing small burst of movement**), fix the annotations, and provide an explanation for the fix.

further improving to 77.1% exact match and 93.6% F1. These methods also resulted in substantially more correct fixes per day (108.7–125.9 min/day) and a moderate increase in incorrect fixes (15.2–27.3 min/day), while the average unchanged time decreased compared to both original annotations and RAG corrections. These findings highlight GLOSS’ ability to go beyond incremental corrections, offering a more comprehensive refinement of activity annotations while maintaining higher agreement with ground truth.

**6.1.2 Qualitative findings.** We present several examples illustrating how GLOSS4HAR integrates multiple sources of sensing data, compares them with participants’ annotations, applied corrections when necessary, and provides a rationale for each adjustment.

First, GLOSS4HAR addresses sloppiness in participant annotations. In Figure 4, we illustrate how in the ACAI dataset, a participant mistakenly merged a short walking episode (about 2–3 minutes, 12:12–12:14) into a longer sitting and eating segment (12:05–12:18), even though the step count data clearly indicated sustained movement. GLOSS4HAR detected this mismatch, inserted a “standing, walking” label, and adjusted the boundaries of the surrounding “sitting and eating” label accordingly. Similarly, in Figure 5, we highlight how GLOSS4HAR identified an overestimation of the walking label (by about 7–8 minutes) by cross-checking the step counts from the Pixel Watch. Additionally, GLOSS4HAR generates justifications for its corrections. These justifications can be displayed in an annotation interface to help researchers or participants quickly understand the reasoning behind each adjustment and decide whether to accept or reject it.

In another example (Figure 6), we illustrate how GLOSS4HAR can detect an omission error. The participant correctly annotated the posture (“standing”) and the primary activity (“cooking”) but forgot to include the secondary activity (“using phone”). By triangulating multiple data sources (in this case, phone usage and ambient sounds) GLOSS4HAR identified the omission, inserted the “using phone” label, and provided a clear justification for the correction. We also highlight another instance where GLOSS4HAR detected a short burst of movement (“walking” from 11:42 to 11:45) and added the corresponding label to the timeline.

## 6.2 GLOSS4HAR potential in supporting low-effort activity labeling (RQ2)

Our results from the timeline reconstruction task show the feasibility of using LLMs to triangulate passive sensing data with participants’ different modes of self-report to generate a timeline of participants’ daily activities. We show the agreement between the timeline generated by GLOSS4HAR and the locomotion detected by the thigh

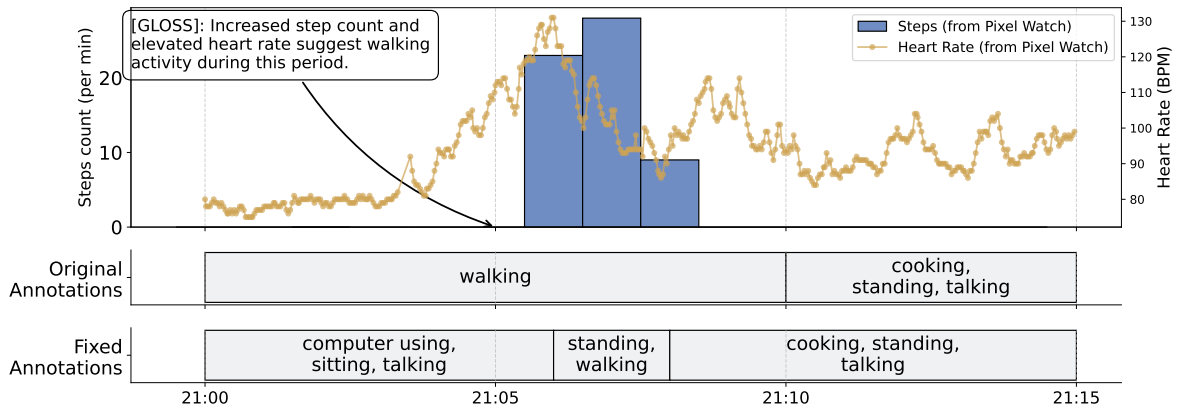


Fig. 5. An example of how GLOSS4HAR can spot a *overflow* error (participants **overestimate the duration** of a label), fix the annotations, and provide an explanation for the fix.

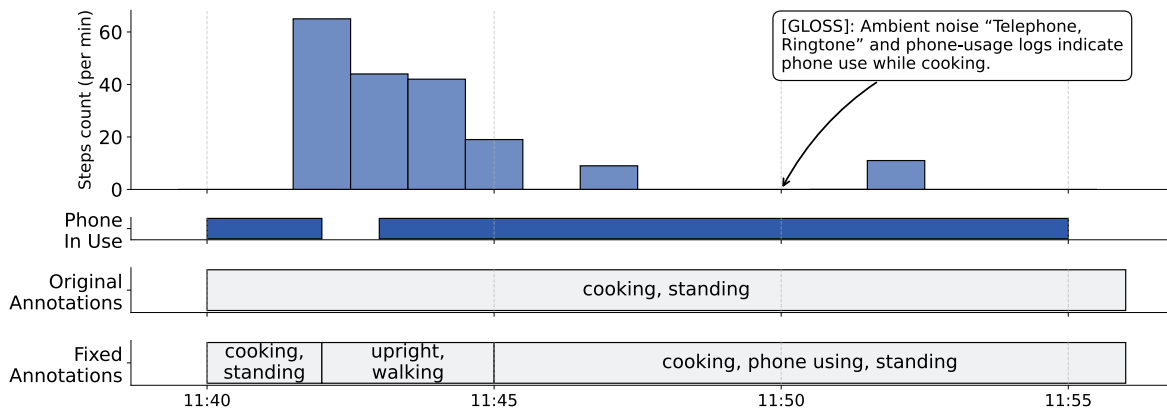


Fig. 6. An example of how GLOSS4HAR can spot an *omission* error (**missing activity label**), based on the evidence verified through multiple data streams (ambient noises and phone logs). In the second subplot (phone in use), **dark blue segments** indicate when the phone is being used (interactive), and white segments indicate the phone is not in use.

sensor, the participant-annotated ground-truth timeline, and the researcher-annotated ground-truth timeline in Table 5.

All four annotation methods show a modest difference in performance when compared against the thigh sensor ground truth, with exact match (EM, the percentage of rows where the predicted labels exactly match the ground truth) ranging from 90.0–98.3%. We observed larger differences while evaluating against researcher and participant annotations. EM, F1 and Jaccard Index (JI, which measures partial overlap between predicted and true labels) highlight these differences. Among the methods, list-based annotation (with approx time and duration) achieved the strongest alignment, with EM of 65.1% (researcher) and 60.9% (participant) and JI of 81.4% and 80.0%. In contrast, the  $\mu$ EMA condition was only marginally higher than the baseline (no self-report), with EM around 44.0–32.7% and JI 71.2–64.9%.

Table 5. Comparison of model performance across self-report conditions against three ground truths: thigh sensor, researcher annotations, and participant annotations. The Thigh Sensor prediction only includes five locomotion (sitting, standing, walking, running, lying down), whereas Researcher and Participant GT use activity labels (Table 2). Metrics include exact match (EM), F1 score (F1), and Jaccard index (JI). Differences in label sets should be considered when interpreting results.

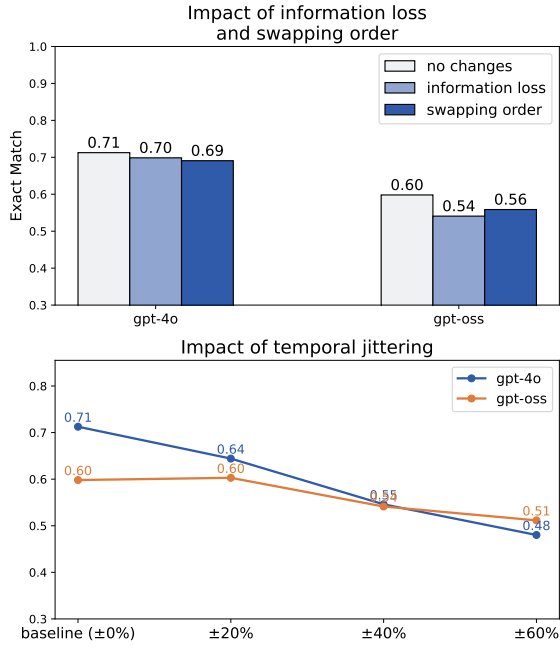
Condition	Model	Thigh Sensor	Researcher GT			Participant GT		
		EM(%)	EM(%)	F1(%)	JI(%)	EM(%)	F1(%)	JI(%)
No self-report	gpt-4o	90.7	33.7	75.5	65.8	25.7	75.2	61.8
	gpt-oss	90.0	37.3	73.5	67.2	32.8	72.6	64.6
$\mu$ EMA	gpt-4o	90.0	44.0	78.9	71.2	32.7	74.4	64.9
	gpt-oss	94.2	37.7	76.7	68.0	29.4	75.5	63.2
List of activities (approx time)	gpt-4o	98.3	65.1	86.7	81.4	60.9	86.4	80.0
	gpt-oss	98.0	51.0	81.1	74.2	45.4	79.8	71.1
List of activities (no time)	gpt-4o	94.8	47.0	81.2	73.2	44.6	81.3	71.2
	gpt-oss	92.9	44.7	78.4	70.9	40.2	77.9	68.5

In conditions such as no self-report or  $\mu$ EMA, there is a substantial gap between EM and JI, indicating that while models often captured some of the correct activities (moderate JI), they rarely predicted the full set exactly (lower EM). In the list-of-activities (no time) condition, EM scores are lower (47.0% researcher, 44.6% participant) compared to the timed list, suggesting that some level of timestamp information is still beneficial, although high temporal precision is not strictly required.

It is interesting to observe this variability in performance across the various self-report conditions. GLOSS4HAR has a strong alignment with the thigh-sensor. We argue this is because the thigh sensor is primarily measuring a limited set of ambulation-based activities like lying down, sitting, standing, walking, and running, that are already quite accurate to measure based on the sensor data without any self-report. Given the already high performance due to the objectiveness and limited set of annotations for the thigh sensor adding additional self-report mechanisms leads to a decent but not a drastic change in performance. The lower performances for the other metrics (especially exact match), however, indicates the need for richer contextual information about participants' routines. We noticed that in the  $\mu$ EMA condition, participants often reported only postures or micro-activities, which contributed little information about the high-level richness of daily life activities. The list-of-activities condition, despite having more temporal imprecision, contains full information about the high-level behaviors of the participants, helping GLOSS4HAR to better narrow down the activities that the participants might have been doing.

**6.2.1 Impact of Information Loss, Activity Swapping, and Temporal Jitter on Reconstruction.** Although the list-of-activities condition introduces multiple sources of imprecision in both labels and timestamps, it still outperforms the no-self-report and  $\mu$ EMA baselines. To better understand which type of imprecision has the greatest impact on reconstruction, we conducted an ablation study examining three factors: information loss (removing or modifying labels), activity swapping (reordering consecutive activities), and temporal jitter (perturbing activity durations). The baseline uses exact labels, exact durations, and correct chronological order from the participants' original annotations. Each ablation condition modifies only one factor from the baseline: information loss removes posture labels and renames activities; activity swapping randomly swaps 20% of consecutive activity pairs; and temporal jitter **increases or decreases all activity's duration** by 20%, 40%, and 60%.

The results (against researcher-corrected ground truths) are shown in Figure 7. Information loss and activity swapping both cause only minor drops in all metrics (EM, F1 and JI), suggesting the reconstruction is robust to



Condition	gpt-oss			gpt-4o		
	EM	F1	JI	EM	F1	JI
<i>(a) Information loss and activity swapping</i>						
Baseline	0.60	0.87	0.80	0.71	0.90	0.85
Info loss	0.54	0.84	0.76	0.70	0.90	0.84
Swapping	0.56	0.85	0.78	0.69	0.90	0.85
<i>(b) Temporal jitter</i>						
Baseline ( $\pm 0\%$ )	0.60	0.87	0.80	0.71	0.90	0.85
$\pm 20\%$	0.60	0.87	0.80	0.64	0.87	0.82
$\pm 40\%$	0.54	0.84	0.77	0.55	0.84	0.77
$\pm 60\%$	0.51	0.84	0.75	0.48	0.81	0.73

Fig. 7. Results of ablation experiments, against researcher-corrected GT. **Left:** Plot of Exact Match between conditions. **Right:** Table of detailed quantitative results between conditions. Information loss and activity swapping introduce relatively small degradation, while temporal jitter causes larger performance drops as perturbation increases.

semantic noises. Temporal jitter, however, has a larger effect when the perturbation becomes substantial ( $\geq 40\%$ ), with EM degrading progressively as the amount of jittering increases.

## 7 TOWARDS REAL-TIME HUMAN-IN-THE-LOOP ACTIVITY RECOGNITION

Our results highlight two main insights. First, GLOSS4HAR can refine activity annotations by cross-checking them with passive sensing data to detect inconsistencies. Second, it can combine passive sensing with low-effort self-reports to construct richer activity timelines. Together, these capabilities suggest the potential for more dynamic, low-effort annotation systems. In particular, our finding that open-ended  $\mu$ EMA fails to capture sufficient activity context – with 59.5% of responses containing only posture information – motivates a shift toward more targeted in-the-moment prompting. Meanwhile, the rich contextual information conveyed in natural language descriptions of daily activities points to narration-based end-of-day reflection as a complementary direction. We propose and evaluate two such systems below.

### 7.1 Context-aware $\mu$ EMA

GLOSS4HAR predicts posture accurately even without self-reports, based on high alignment with thigh sensor data ( $\approx 90\%$  accuracy). The bottleneck, therefore, is not posture but high-level activity context. This suggests that  $\mu$ EMA questions should be redesigned to constrain responses in ways that help narrow down possible activities, rather than asking open-ended questions that participants tend to answer with posture alone.

We propose a real-time, human-in-the-loop annotation system (Figure 8) that operates through three processes:

- **$\mu$ EMA generation:** At 15-minute intervals, GLOSS4HAR triangulates passive sensing data to generate a list of likely activities and prompts participants with context-aware, personalized yes/no questions targeting the most probable activities.
- **Activity annotation:** Once participants respond, GLOSS4HAR integrates the  $\mu$ EMA responses with passive sensing data to produce updated activity predictions and extend the timeline.
- **Retrospective correction:** When new self-reports or sensing data become available, GLOSS4HAR revisits prior predictions and updates the activity timeline accordingly.

### 7.2 Narration-based End-of-day Reflection

A popular approach for retrospective recall is *narration-based annotation*, where participants describe their day in natural language at the end of the day – for instance, through voice or text input on a smartphone or smartwatch. This approach is flexible and aligns with how people naturally recall their day. Unlike list-based methods, narration requires less structured thinking from participants and can capture both temporal and categorical uncertainty (e.g., “I started walking to campus *sometime between 8:00 am and 8:05 am*” vs. “I sat and used a computer from 1–3 pm, *with a bathroom break in between*”). This makes it potentially well-suited for longitudinal studies where minimizing participant burden is critical.

### 7.3 Experiment Design

**7.3.1 Context-aware  $\mu$ EMA simulation.** We conducted a simulation study using the ACAI dataset to evaluate how GLOSS4HAR would function in a real-time, human-in-the-loop setting. We replicated the  $\mu$ EMA schedule (every 15 minutes) and participant non-response rate observed in the ACAI study. At every  $\mu$ EMA time point, GLOSS4HAR was prompted with the current timestamp, past  $\mu$ EMA responses, and the generated activity timeline so far. GLOSS4HAR responded with an updated timeline and up to three most likely activities for the current segment. We simulated the next  $\mu$ EMA question as a yes/no prompt (e.g., “Are you doing [X]?”), with participant responses simulated based on researcher-annotated ground truth – answering “Yes” if the true activity matched one of the suggestions and “No” otherwise. We only simulated a response if the participant had actually responded to the standard  $\mu$ EMA in the original ACAI study, preserving ecological validity. This cycle repeated throughout the day, allowing GLOSS4HAR to iteratively update the activity timeline while integrating simulated user feedback.

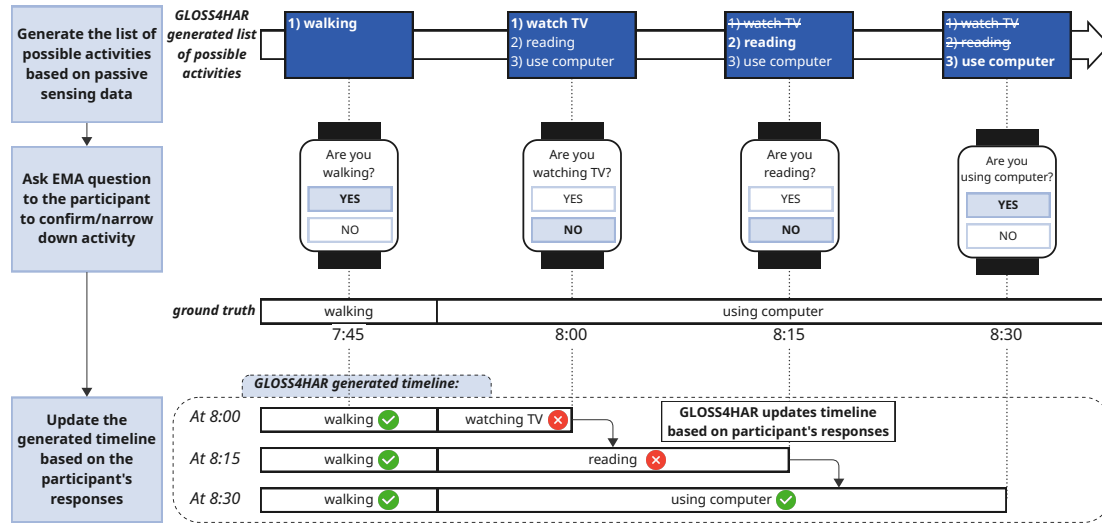


Fig. 8. Our proposed real-time, human-in-the-loop activity annotation system that incorporates both the abilities of GLOSS4HAR to correct and generate activity labels. Using passive sensing data, GLOSS4HAR can suggest what the participant is doing in-the-moment and generate personalized  $\mu$ EMA prompts to get the most contextual information to make the predictions. After the participant responds, GLOSS4HAR can update its memory and understanding of the passive sensing data, to either: 1) extend the activity timeline to include prediction up to that point, or 2) correct past annotations based on new information collected. In this example, GLOSS4HAR detects continuous activity from 7:50 to 8:30 am and initially identifies three possible activities: “watching TV,” “reading,” or “using computer.” At 8:00 am and 8:15 am, the participant rejected the first two options in their  $\mu$ EMA responses at 8:00 and 8:15 am. When asked at 8:30 am whether they were “using computer,” the participant confirmed, allowing GLOSS4HAR to label the entire period from 7:50 to 8:30 am as “using computer.”

**7.3.2 Narration simulation.** To explore the potential of narration-based annotation, we simulated free-form participant narrations by prompting an LLM (gpt-4o) to generate natural language descriptions of each participant’s day from their detailed ACAI annotations. These synthetic narrations were then provided to GLOSS4HAR as input for reconstructing activity timelines. An example of a narration is shown in Figure 9. We note an important caveat: because narrations were synthetically generated by an LLM, there is a potential circularity – GLOSS4HAR may be especially well-suited to processing LLM-generated text, which may not reflect the variability, ambiguity, and omissions present in real participant narrations. As such, these results should be interpreted as a proof-of-concept motivating future data collection rather than a validated empirical finding.

## 7.4 Results

Results from both systems are presented in Table 6. The context-aware  $\mu$ EMA approach substantially outperforms all other annotation conditions, with exact match against researcher-annotated ground truth reaching 88.9% with gpt-4o – more than double the standard  $\mu$ EMA condition and 13 percentage points higher than the list-of-activities (approx time) condition. This confirms the viability of context-aware, targeted prompting as a practical direction for future real-time annotation systems.

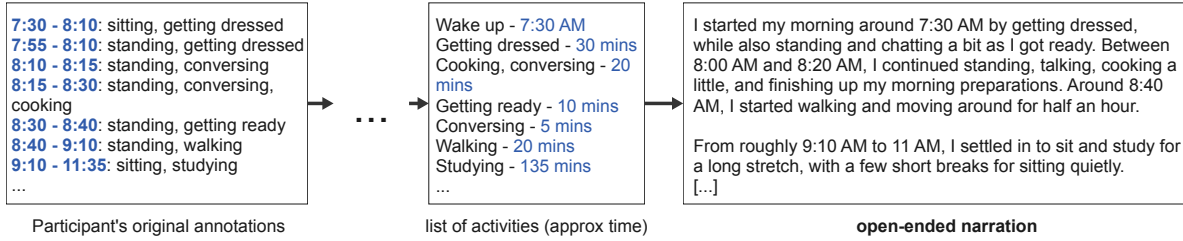


Fig. 9. Simulation of open-ended narration. We passed the list of activities (approx time) self-reports to an LLM (gpt-4o), and prompted it to generate a first-person narration of the entire day.

Table 6. Results of the real-time context-aware  $\mu$ EMA simulation experiment and narration-based annotation, compared to other low-effort annotation methods. Thigh sensor shows Exact Match only.

Condition	Model	Thigh Sensor	Researcher GT			Participant GT		
		EM(%)	EM(%)	F1(%)	J1(%)	EM(%)	F1(%)	J1(%)
No self-report	gpt-4o	90.7	33.7	75.5	65.8	25.7	75.2	61.8
	gpt-oss	90.0	37.3	73.5	67.2	32.8	72.6	64.6
$\mu$ EMA	gpt-4o	90.0	44.0	78.9	71.2	32.7	74.4	64.9
	gpt-oss	94.2	37.7	76.7	68.0	29.4	75.5	63.2
List of activities (approx time)	gpt-4o	98.3	65.1	86.7	81.4	60.9	86.4	80.0
	gpt-oss	98.0	51.0	81.1	74.2	45.4	79.8	71.1
List of activities (no time)	gpt-4o	94.8	47.0	81.2	73.2	44.6	81.3	71.2
	gpt-oss	92.9	44.7	78.4	70.9	40.2	77.9	68.5
Narration	gpt-4o	96.4	75.7	92.0	87.8	73.9	91.5	86.9
	gpt-oss	98.3	59.5	85.6	79.4	57.3	88.6	78.3
Context-aware $\mu$ EMA	gpt-4o	99.7	88.9	95.9	94.1	84.6	94.8	91.9
	gpt-oss	99.2	63.4	85.7	81.1	56.7	85.1	77.5

The narration-based condition achieved the strongest performance among all retrospective annotation conditions (EM: 75.7%, F1: 92.0%, JI: 87.8% against researcher ground truth with gpt-4o), outperforming the list-of-activities (approx time) condition by over 10 percentage points in exact match. Given the caveat about circularity noted above, these results are best understood as an upper bound on what narration-based annotation could achieve if participants provided detailed, coherent end-of-day descriptions. Future work should collect real participant narrations to validate this potential.

## 8 DISCUSSIONS AND FUTURE WORKS

We discuss how our findings extended the existing research on human activity tracking and annotation from wearables data, reflect on the implications for future AI-assisted activity annotation systems, and highlight the potential of balancing passive and active sensing for long-term engagement and retention.

### 8.1 Designing future human-AI collaborative annotation system (RQ1)

Integrating AI or automated assistance into activity annotation is not a new topic in HCI and UbiComp [74, 80]. Prior works have shown that even simple automation—such as segmenting activity start and stop times or estimating durations—can substantially reduce participants’ cognitive burden [6, 80]. Most assisted-recall systems, however, rely on heuristic or machine learning–based prediction algorithms that do not provide human-level explanations of the underlying data or the suggested labels and corrections. Furthermore, most activity annotation systems focus solely on suggesting possible activities, with few offering real-time flagging of potential errors in participants’ annotations.

Our work demonstrates how LLMs can assist researchers in refining and cleaning activity annotations, substantially reducing the time and effort required for dataset curation [46]. We show that an multi-agent LLM system (GLOSS4HAR) can effectively parse passive sensing data, review participants’ annotations, detect potential errors or misalignments, and propose appropriate fixes. Our experiments reveal that GLOSS4HAR’s corrections outperform both participants’ original annotations and RAG-based models in terms of alignment with the thigh sensor and researcher-reviewed ground truth.

Building on these findings, we envision future human–AI collaborative annotation systems that integrate LLMs like GLOSS4HAR directly into the annotation workflow. Such systems could not only assist researchers during post-hoc data cleaning but also provide real-time feedback/explanations to participants during data annotation process, improving annotation quality. Future research should focus on designing these interactive annotation interfaces and examining how LLM-assisted annotation affects participant engagement, effort, and data quality [7, 11].

### 8.2 Balancing the need for participants’ self-report and passive measurements (RQ2)

In longitudinal research studies, researchers must carefully balance the need for participants’ self-reports with passive measurements [22, 75, 101]. Too frequent or too lengthy self-reporting can lead to fatigue, reduced engagement, and ultimately lower data quality and participant retention. Self-reports remain essential because no detection model is perfect—only participants can fully discern the intricacies of their daily routines and provide context for unusual or ambiguous data. For this reason, many studies have explored diverse ways to collect self-reports, using incentives such as monetary rewards [40], gamification [93, 94], or varied survey questions [62] to encourage compliance.

The inclusion of multiple self-report methods can introduce inconsistencies, as not all formats provide the same level of detail [55]. In this work, we demonstrate that by leveraging the open-ended reasoning and commonsense knowledge of LLMs, we can integrate diverse, low-burden self-report formats to reconstruct activity timelines that align closely with both participants’ original and researcher-reviewed annotations. This suggests the potential of LLMs to serve as a mediating layer that reconciles insights across multimodal data sources and communicates them back to both researchers and participants in meaningful ways. Such capability is unique to LLMs and cannot be achieved with traditional ML-based approaches, making this an especially promising direction for future research.

Interestingly, our results show that the most cognitively demanding part of annotation, temporal precision, contributed little to the model’s ability to generate accurate timelines. This finding suggests an opportunity for human–AI collaboration in which LLMs, having access to sensory data and self-reports, handle data segmentation and triangulation, while participants provide concise contextual insights, thereby reducing effort on both sides and improving overall annotation efficiency and quality [91, 98]. Longitudinal studies are needed to examine the long-term effects of LLM-assisted annotation, as such systems have the potential to gradually learn participants’ routines and activities over time, further reducing reliance on participants’ self-reports. Researchers, however,

should be wary of potential troublesome annotation behaviors, such as under- or over-reliance on LLMs for annotation [12, 67, 111].

### 8.3 Integration with real-time HAR system

Recent advances in LLMs have created new opportunities for HAR, from generating natural language explanations for predictions and helping researchers interpret sensor data to processing multimodal, high-frequency sensing streams [16, 19, 63, 110]. Yet, few studies have explored how such systems can support real-time, naturalistic activity tracking [90].

Our results demonstrate the feasibility of using an LLM-based system (GLOSS4HAR) to improve activity annotation quality while reducing both participant and researcher burden. Beyond post-hoc data cleaning, GLOSS4HAR could also support real-time HAR by generating context-aware  $\mu$ EMA prompts tailored to users' current situations [20, 62], enabling richer and more targeted self-reports over time. GLOSS4HAR could also potentially support open-ended form of self-reports, like end-of-day narration style diary, which have been shown to have significantly less burdensome for some population like older adults [78, 99]. More broadly, our findings point toward future HAR systems that move beyond passive sensing to support flexible, human-AI collaboration for physical activities and behavior understanding.

At the same time, privacy remains a key challenge for LLM-assisted HAR systems, as sending raw sensor data, self-reports, or contextual information to cloud APIs may undermine user trust and data security. To address this, we used a local open-source LLM so sensitive data could remain within the local server. This privacy-preserving setup, however, came with lower performance than larger proprietary models. Future work could improve local deployment by fine-tuning smaller, specialized sensor-aware models, such as some recent works in LLM-sensor aligned models [38, 63, 102, 108, 110], to better support annotation correction and timeline reconstruction while maintaining strong privacy guarantees.

## 9 LIMITATIONS

We demonstrated the potential of GLOSS4HAR for annotation correction and low-effort data collection, but several limitations remain. First, we use researcher annotations as ground truth. While not absolute (e.g., compared to egocentric video), they represent the most reliable reference under real-world constraints without direct observation. This choice aligns with our goal of reducing researcher burden, though future work should validate against visually derived ground truth.

Second, the system can have high latency, especially with GPT-4o, where responses may take up to a minute due to API and network dependencies, limiting real-time use. Sending sensing data to external APIs also raises privacy concerns. We include a smaller local model (gpt-oss:20b) to mitigate this, though it currently underperforms; improvements in local LLMs may address this gap.

Third, the ACAI dataset includes only phone and smartwatch data, which may limit generalizability to settings with different sensor configurations (e.g., smart homes). In addition, our timeline generation relies on simulated self-reports derived from detailed annotations, which may not fully capture real-world variability or omissions. Future work should evaluate performance with real self-reports.

Finally, we did not conduct a user study to assess effects on annotator behavior or engagement. As a feasibility study, we focused on technical viability; future work should examine real-world deployment and user experience.

## 10 CONCLUSION

Our work demonstrates the feasibility of LLMs as powerful collaborators in the future of human activity tracking and recognition. We show that LLMs can move beyond traditional classification tasks to perform sensemaking—reconciling multiple data streams, identifying inconsistencies, and constructing coherent activity timelines

from minimal participant input. Our findings highlight the potential of LLMs to bridge the gap between passive sensing and active self-report, transforming fragmented, multimodal data into structured insights that more accurately reflect human behavior. Looking ahead, we envision LLM-driven systems that can continuously interpret and synthesize diverse streams of sensing data, supporting more adaptive, context-aware, and personalized insights and interventions that evolve alongside individuals over time.

## REFERENCES

- [1] Rebecca Adaimi and Edison Thomaz. Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–23, September 2019.
- [2] Rebecca Adaimi and Edison Thomaz. Lifelong adaptive machine learning for sensor-based human activity recognition using prototypical networks. *Sensors*, 22(18):6881, September 2022.
- [3] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. PrISM-Q&A: Step-aware voice assistant on a smartwatch enabled by multimodal procedure tracking and large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–26, November 2024.
- [4] Riku Arakawa, Hiromu Yakura, and Mayank Goel. PrISM-Observer: Intervention agent to help users perform everyday procedures sensed using a smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16. ACM, October 2024.
- [5] Riku Arakawa, Hiromu Yakura, and Masataka Goto. BeParrot: Efficient interface for transcribing unclear speech via respeaking. In *27th International Conference on Intelligent User Interfaces*, pages 832–840, Helsinki Finland, March 2022. ACM.
- [6] Anonymized Authors. A context-assisted, semi-automated activity recall interface allowing uncertainty. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):33, 2025.
- [7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, May 2021.
- [8] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In 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, editors, *Pervasive Computing*, volume 3001, pages 1–17. Springer Berlin Heidelberg, 2004.
- [9] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, Kemal Eren, Jaime I Cervantes, Buote Xu, Fynn Beuttenmueller, Adrian Wolny, Chong Zhang, Ullrich Koethe, Fred A. Hamprecht, and Anna Kreshuk. ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods*, 16(12):1226–1232, December 2019.
- [10] Luigi Bibbò and Artur Serrano. Human activity recognition (HAR) in healthcare, 2nd Edition. *Applied Sciences*, 15(10):5762, May 2025.
- [11] Claus Bossen and Kathleen H. Pine. Batman and Robin in healthcare knowledge work: Human-AI collaboration by clinical documentation integrity specialists. *ACM Transactions on Computer-Human Interaction*, page 3569892, October 2022.
- [12] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, April 2021.
- [13] Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data*, 11(1):1135, October 2024.
- [14] Yung-Ju Chang, Gaurav Paruthi, and Mark W. Newman. A field study comparing approaches to collecting annotated activity data in real-world settings. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 671–682. ACM, September 2015.
- [15] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. The Opportunity Challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, November 2013.
- [16] Wenqiang Chen, Jiaxuan Cheng, Leyao Wang, Wei Zhao, and Wojciech Matusik. 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):1–26, November 2024.
- [17] Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tiejun Qian, and Minlie Huang. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194. Association for Computational

- Linguistics, 2024.
- [18] Akshat Choube, Sohini Bhattacharya, Rahul Majethia, Jiachen Li, Vedant Das Swain, and Varun Mishra. Imputation matters: A deeper look into an overlooked step in longitudinal health and behavior sensing research. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):1–30, December 2025.
  - [19] Akshat Choube, Ha Le, Jiachen Li, Kaixin Ji, Vedant Das Swain, and Varun Mishra. GLOSS: Group of LLMs for open-ended sensemaking of passive sensing data for health and wellbeing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):30, 2025.
  - [20] Akshat Choube, Vedant Das Swain, and Varun Mishra. SeSaMe: A framework to simulate self-reported ground truth for mental health sensing studies, 2024.
  - [21] Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065, December 2021.
  - [22] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. Semantic gap in predicting mental wellbeing through passive sensing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, April 2022.
  - [23] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimjoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. Increasing the speed and accuracy of data labeling through an AI assisted interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401. ACM, April 2021.
  - [24] Aiden R. Doherty, Niamh Caprani, Ciarán Ó Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F. Smeaton, and Noel E. O’Connor. Passively recognising human activities through lifelogging. *Computers in Human Behavior*, 27(5):1948–1958, September 2011.
  - [25] Genevieve Fridlund Dunton, Eldin Dzubur, Keito Kawabata, Brenda Yanez, Bin Bo, and Stephen Intille. Development of a smartphone application to measure physical activity using sensor-assisted self-report. *Frontiers in Public Health*, 2:12, 2014.
  - [26] Sara Evensen, Chang Ge, and Cagatay Demiralp. Ruler: Data programming by demonstration for document labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1996–2005. Association for Computational Linguistics, 2020.
  - [27] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. PhysioLLM: Supporting personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8, Houston, TX, USA, November 2024. IEEE.
  - [28] Google, Inc. Google Cloud Speech-to-text, 2026.
  - [29] Google Research. YAMNet: Yet another multilabel neural network for audio event recognition., 2017.
  - [30] Moritz A. Graule and Volkan Isler. GG-LLM: Geometrically Grounding Large Language Models for Zero-shot Human Activity Forecasting in Human-Aware Task Planning, October 2023. arXiv:2310.20034 [cs].
  - [31] Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990. IEEE, June 2022.
  - [32] Harish Haresamudram, Apoorva Beedu, Mashfiqui Rabbi, Sankalita Saha, Irfan Essa, and Thomas Ploetz. Limitations in employing natural language supervision for sensor-based human activity recognition - and ways to overcome them. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):273–281, April 2025.
  - [33] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables, February 2022. arXiv:2202.12938 [cs, eess].
  - [34] Paul R. Hibbing, Nicholas R. Lamoureux, Charles E. Matthews, and Gregory J. Welk. Protocol and data description: The free-living activity study for health. *Journal for the Measurement of Physical Behaviour*, 4(3):197–204, September 2021.
  - [35] Michael J Hoefer, Raegan Rychecky, Max Gong, and Stephen Volda. TellTime: An AI-augmented calendar with a voice interface for collecting time-use data. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1366–1380. ACM, March 2025.
  - [36] Alexander Hoelzemann and Kristof Van Laerhoven. A matter of annotation: An empirical study on in situ and self-recall activity annotations from wearable sensors. *Frontiers in Computer Science*, 6:1379788, July 2024.
  - [37] Xin Hu, Dianhui Chu, Zhengzuo Li, Ruixuan Dai, Yang Cui, Zhiyuan Zhou, Boyang An, Yongqing Han, Chunmao Jiang, and Deqiong Ding. Coarse-to-fine activity annotation and recognition algorithm for solitary older adults. *IEEE Access*, 8:4051–4064, 2020.
  - [38] Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. LLaSA: A sensor-aware LLM for natural language reasoning of human activity from IMU data, September 2025. arXiv:2406.14498 [cs].
  - [39] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. micro-EMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1124–1128. ACM, September 2016.
  - [40] Sueun Jang, Youngseok Seo, Woohyeok Choi, and Uichin Lee. Like adding a small weight to a scale about to tip: Personalizing micro-financial incentives for digital wellbeing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, April 2025.

- [41] Yugyeong Jung, Hei Yiu Law, Hadong Lee, Junmo Lee, Bongshin Lee, and Uichin Lee. DataSentry: Building missing data management system for In-the-wild mobile sensor data collection through multi-year iterative design approach. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18. ACM, April 2025.
- [42] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM, April 2020.
- [43] Sarah Kozey Keadle, Shreya Patel, David Berrigan, Cami N. Christopher, Jeffery Huang, Pedro F. Saint-Maurice, Erikka Loftfield, and Charles E. Matthews. Validation of ACT24 version 2.0 for estimating behavioral domains, active and sedentary time. *Medicine & Science in Sports & Exercise*, 55(6):1054–1062, June 2023.
- [44] Paul Kelly, Simon J. Marshall, Hannah Badland, Jacqueline Kerr, Melody Oliver, Aiden R. Doherty, and Charlie Foster. An ethical framework for automated, wearable cameras in health behavior research. *American Journal of Preventive Medicine*, 44(3):314–319, March 2013.
- [45] Jacqueline Kerr, Simon J. Marshall, Suneeta Godbole, Jacqueline Chen, Amanda Legge, Aiden R. Doherty, Paul Kelly, Melody Oliver, Hannah M. Badland, and Charlie Foster. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. *American Journal of Preventive Medicine*, 44(3):290–296, March 2013.
- [46] Hossein Khayami, Lining Wang, Young-Ho Kim, Bongshin Lee, David E. Conroy, Amanda Lazar, Eun Kyoung Choe, and Hernisa Kacorri. From verbal reports to personalized activity trackers: Understanding the challenges of ground truth data collection with older adults in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):1–33, June 2025.
- [47] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. MyMove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *CHI Conference on Human Factors in Computing Systems*, pages 1–21, New York, NY, USA, April 2022. ACM.
- [48] Youngwon Kim and Gregory J. Welk. The accuracy of the 24-h activity recall method for assessing sedentary behaviour: The physical activity measurement survey (PAMS) project. *Journal of Sports Sciences*, 35(3):255–261, February 2017.
- [49] Simone Kohler, Gundula Behrens, Matthias Olden, Sebastian E Baumeister, Alexander Horsch, Beate Fischer, and Michael F Leitzmann. Design and evaluation of a computer-based 24-hour physical activity recall (cpar24) instrument. *Journal of Medical Internet Research*, 19(5):e186, May 2017.
- [50] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, March 2011.
- [51] Hyeokhyen Kwon, Gregory D. Abowd, and Thomas Plötz. Handling annotation uncertainty in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 109–117. ACM, September 2019.
- [52] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38. ACM, January 2019.
- [53] Rithika Lakshminarayanan, Arushi Uppal, Ha Le, James Spilsbury, and Stephen Intille. Detecting sleep disruptions in adolescents using context-sensitive ecological momentary assessment: A feasibility study. In *Proceedings of the 18th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 1–12, New York, NY, USA, 2024. ACM.
- [54] Ha Le, Rithika Lakshminarayanan, Jixin Li, Varun Mishra, and Stephen Intille. Collecting self-reported physical activity and posture data using audio-based ecological momentary assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–35, 2024.
- [55] Ha Le, Veronika Potter, Rithika Lakshminarayanan, Varun Mishra, and Stephen Intille. Feasibility and utility of multimodal micro ecological momentary assessment on a smartwatch. *CHI Conference on Human Factors in Computing Systems (CHI’ 25)*, 2025.
- [56] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. EchoWrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21. ACM, May 2024.
- [57] Jong Ho Lee, Sunghoon Ivan Lee, and Eun Kyoung Choe. GoalTrack: Supporting Personalized Goal-Setting in Stroke Rehabilitation with Multimodal Activity Journaling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, November 2024.
- [58] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. IMUGPT 2.0: Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):32, August 2024.
- [59] 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, pages 9459–9474, Red Hook, NY, USA, December 2020. Curran Associates Inc.

- [60] Huichuwu Li, Jiang Xiao, Wei Wang, Lu Wang, Dian Zhang, and Hai Jin. InFit: Combination movement recognition for intensive fitness assistant via Wi-Fi. *IEEE Transactions on Mobile Computing*, pages 1–15, 2022.
- [61] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. Vital Insight: Assisting experts’ context-driven sensemaking of multi-modal personal tracking data using visualization and human-in-the-loop LLM agents, 2024.
- [62] Jixin Li, Aditya Ponnada, Wei-Lin Wang, Genevieve Dunton, and Stephen Intille. Ask less, learn more: Adapting ecological momentary assessment survey length by modeling question-answer information gain. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–32, November 2024.
- [63] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Human-intuitive alignment of multivariate sensor data with LLMs for activity recognition, May 2025.
- [64] Hyunchul Lim, Nam Anh Dang, Dylan Lee, Tianhong Catherine Yu, Jane Lu, Franklin Mingzhe Li, Yiqi Jin, Yan Ma, Xiaojun Bi, François Guimbretière, and Cheng Zhang. SpellRing: Recognizing continuous fingerspelling in American sign language using a ring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17. ACM, April 2025.
- [65] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kumpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–28, December 2020.
- [66] Yuhan Luo, Bongshin Lee, Young-Ho Kim, and Eun Kyoung Choe. NoteWordy: Investigating touch and speech input on smartphones for personal data capture. *Proceedings of the ACM on Human-Computer Interaction*, 6(ISS):568–591, November 2022.
- [67] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, April 2023.
- [68] Charles E. Matthews. Twenty-four hour physical activity recall (24PAR) system interviewer training materials/protocol.
- [69] Mike A. Merrill, Akshay Paruchuri, Naghme Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. Transforming wearable data into personal health insights using large language model agents, 2024.
- [70] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. Investigating contextual cues as indicators for EMA delivery. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 935–940. ACM, September 2017.
- [71] Naoya Miyake, Haru Kaneko, Elsen Ronando, Christina Garcia, and Sozo Inoue. Toward detecting and explaining stress of nurses using wearable devices and LLMs. In José Bravo, Chris Nugent, and Ian Cleland, editors, *Proceedings of the International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2024)*, volume 1212, pages 288–299. Springer Nature Switzerland, Cham, 2024. Series Title: Lecture Notes in Networks and Systems.
- [72] Subigy Nepal, Arvind Pillai, William Campbell, Talie Massachi, Eunsol Soul Choi, Orson Xu, Joanna Kuc, Jeremy Huckins, Jason Holden, Colin Depp, Nicholas Jacobson, Mary Czerwinski, Eric Granholm, and Andrew T. Campbell. Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App, March 2024. arXiv:2404.00487 [cs].
- [73] Subigy Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. MindScape study: Integrating LLM and behavioral sensing for personalized AI-Driven journaling experiences. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–44, November 2024.
- [74] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M. Almeida, and Santosh Kumar. Momentary stressor logging and reflective visualizations: Implications for stress management with wearables. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, May 2024.
- [75] Aditya Ponnada. Longitudinal user engagement with microinteraction ecological momentary assessment ( $\mu$ EMA). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):27, 2025.
- [76] Aditya Ponnada, Jixin Li, Shirlene Wang, Wei-Lin Wang, Bridgette Do, Genevieve F. Dunton, and Stephen S. Intille. Contextual biases in microinteraction ecological momentary assessment (micro-EMA) non-response. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–24, March 2022.
- [77] Veronika Potter, Hoan Tran, Daniel Mobely, Suzanne M. Bertisch, Dinesh John, and Stephen Intille. The Physical Activity Assessment Using Wearable Sensors (PAAWS) Dataset: Labeled Laboratory and Free-Living Accelerometer Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):32, December 2025.
- [78] Alisha Pradhan, Amanda Lazar, and Leah Findlater. Use of intelligentvoice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction*, 27(4):1–27, August 2020.

- [79] Mashfiqui Rabbi, Katherine Li, H. Yanna Yan, Kelly Hall, Predrag Klasnja, and Susan Murphy. ReVibe: A context-assisted evening recall approach to improve self-report adherence. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–27, December 2019.
- [80] Mashfiqui Rabbi, Angela Pfammatter, Mi Zhang, Bonnie Spring, and Tanzeem Choudhury. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth and uHealth*, 3(2):e42, May 2015.
- [81] Akane Sano, Judith Amores, and Mary Czerwinski. Exploration of LLMs, EEG, and behavioral data to measure and support attention and sleep, 2024.
- [82] Lisa Schrader, Agustín Vargas Toro, Sebastian Konietzny, Stefan Rüping, Barbara Schäpers, Martina Steinböck, Carmen Krewer, Friedemann Müller, Jörg Güttler, and Thomas Bock. Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing*, 13(2):139–165, June 2020.
- [83] Nazanin Sedaghati, Sondas Ardebili, and Ali Ghaffari. Application of human activity/action recognition: a review. *Multimedia Tools and Applications*, 84(28):33475–33504, January 2025.
- [84] Abigail J. Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. Do life-logging technologies support memory for the past?: An experimental study using sensecam. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 81–90. ACM, April 2007.
- [85] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4(1):1–32, April 2008.
- [86] Sloke Shrestha and Edison Thomaz. Leveraging large language models to annotate activities of daily living captured with egocentric vision. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 182–186. IEEE, June 2024.
- [87] Georgios Sopidis, Michael Haslgrübler, Behrooz Azadi, Bernhard Anzengruber-Tánase, Abdelrahman Ahmad, Alois Ferscha, and Martin Baresch. Micro-activity recognition in industrial assembly process with IMU data and deep learning. In *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments*, pages 103–112. ACM, June 2022.
- [88] Marija Stojchevska, Mathias De Brouwer, Martijn Courteaux, Femke Ongenaes, and Sofie Van Hoecke. From lab to real world: Assessing the effectiveness of human activity recognition and optimization through personalization. *Sensors*, 23(10):4606, May 2023.
- [89] Lie Ming Tang, Jochen Meyer, Daniel A. Epstein, Kevin Bragg, Lina Engelen, Adrian Bauman, and Judy Kay. Defining adherence: Making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–22, March 2018.
- [90] Megha Thukral, Sourish Gunesh Dhekane, Shruthi K. Hiremath, Harish Haresamudram, and Thomas Ploetz. Layout-agnostic human activity recognition in smart homes through textual descriptions of sensor triggers (TDOST). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1):1–38, March 2025.
- [91] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, October 2024.
- [92] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. ExtraSensory App: Data collection in-the-wild with rich user interface to self-report behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, April 2018.
- [93] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–21, September 2017.
- [94] Niels Van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, 134:1–12, February 2020.
- [95] Andrew Vargo, Shoya Ishimaru, Md. Rabiul Islam, Benjamin Tag, and Koichi Kise. Obtaining labels for in-the-wild studies: Using visual cues and recall. *IEEE Pervasive Computing*, 21(1):9–17, January 2022.
- [96] Igor Vurdelja, Elena Vučeljić, Jelena Medarević, Marko Marković, Milica Janković, and Zaharije Radivojević. Smart body sensor network for logging of activities of daily living. In *7th Conference on the Engineering of Computer Based Systems*, pages 1–10. ACM, May 2021.
- [97] Chongyang Wang, Yuan Feng, Lingxiao Zhong, Siyi Zhu, Chi Zhang, Siqi Zheng, Chen Liang, Yuntao Wang, Chengqi He, Chun Yu, and Yuanchun Shi. UbiPhysio: Support daily functioning, fitness, and rehabilitation with action understanding and feedback in natural language. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–27, March 2024.
- [98] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM, April 2020.
- [99] Yiwen Wang, Mengying Li, Young-Ho Kim, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. Redefining activity tracking through older adults’ reflections on meaningful activities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM, May 2024.

- [100] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology*, 2(1):1–23, January 2011.
- [101] Cornelia Wrzus and Andreas B. Neubauer. Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, 30(3):825–846, April 2023.
- [102] Wenxuan Xu, Arvind Pillai, Subigya Nepal, Amanda C. Collins, Daniel M. Mackin, Michael V. Heinz, Tess Z. Griffin, Nicholas C. Jacobson, and Andrew Campbell. LENS: LLM-enabled narrative synthesis for mental health by aligning multimodal sensing with language models, December 2025. arXiv:2512.23025 [cs].
- [103] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, December 2022.
- [104] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, March 2024.
- [105] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, November 2024.
- [106] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):1–5, January 2025.
- [107] Xiaofan Yu. SensorChat: Answering qualitative and quantitative questions during long-term multimodal sensor interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3), 2025.
- [108] Aiwei Zhang, Arvind Pillai, Andrew Campbell, and Nicholas C Jacobson. MotionTeller: Multi-modal integration of wearable time-series with LLMs for health and behavioral understanding.
- [109] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. EchoSpeech: Continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18. ACM, April 2023.
- [110] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the Language of Wearable Sensors, 2025.
- [111] Jiawei Zhou. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. 2023.

## A APPENDIX A: PROMPT FOR GENERATING ACTIVITY NARRATION

To generate the open-ended narration of activities, we used the self-reports generated from the list of activities (approx time) condition, and ask gpt-4o to write a paragraph about what a person would have said about their day. This is the prompt we used to pass to the LLM:

This is the list of activities an individual has done throughout the day, with approximate timestamp or duration.

[insert the self-reports used in list of activities (approx time) condition]

Simulate what it would be like if instead of a list of activities, the person would speak or write a paragraph of what happened during their day (chronologically, with a mix of activity duration and timestamp). Generate the paragraph using the data above.

## B APPENDIX B: AN EXAMPLE OF HOW GLOSS4HAR WORKS FOR ANNOTATION CORRECTION TASK

**User Query:** Subject with id pilot7 annotated sitting, attending meeting during the entire period from 13:23–14:29 on 2025/02/27.

Based on passive sensing data, provide more accurate annotations:

- Check 10 minutes before and after the interval to refine start/end times.
- Correct posture and activities if needed.
- If the participant did multiple things, include all relevant activities.
- Merge consecutive identical labels.
- Do not output overlapping annotations.

### Action Plan Generation Agent – Action Plan:

- (1) Define target interval: 13:23–14:29 and expand to 13:13–14:39 for context.
- (2) Query step count data to identify movement and stationary periods.
- (3) Check watch wear status to ensure data reliability.
- (4) Query heart rate trends for signs of exertion or sustained sedentary activity.
- (5) Query phone usage logs to detect active phone interaction.
- (6) Query ambient noise labels every 5 minutes for speech or environmental cues.
- (7) Query GPS location records for mobility and location context.
- (8) Triangulate all signals to refine posture and activities.
- (9) Merge consecutive segments with identical labels.

### Coding and Database Management Agents – Retrieved Sensor Data (13:13–14:39):

```
{
  "step_count": [
    {"start_time": "13:21:46", "end_time": "13:21:46", "avg_steps_per_min": 10.0},
    {"start_time": "13:24:13", "end_time": "13:26:32", "avg_steps_per_min": 58.71},
    {"start_time": "13:33:16", "end_time": "13:33:16", "avg_steps_per_min": 6.0},
    {"start_time": "13:37:56", "end_time": "13:37:56", "avg_steps_per_min": 6.0},
    {"start_time": "13:42:29", "end_time": "13:42:29", "avg_steps_per_min": 4.0},
    {"start_time": "13:50:24", "end_time": "13:50:24", "avg_steps_per_min": 8.0},
    {"start_time": "14:01:29", "end_time": "14:01:29", "avg_steps_per_min": 4.0},
    {"start_time": "14:22:32", "end_time": "14:22:32", "avg_steps_per_min": 1.0},
    {"start_time": "14:32:01", "end_time": "14:32:01", "avg_steps_per_min": 9.0}
  ],
  "watch_non_wear": [],
  "heart_rate": [
    {"start_time": "13:13:00", "end_time": "13:22:25", "avg_hr": 83.4},
    {"start_time": "13:22:27", "end_time": "13:54:25", "avg_hr": 83.0},
    {"start_time": "13:54:27", "end_time": "14:38:59", "avg_hr": 77.7}
  ],
  "phone_usage": [],
  "ambient_noise": [
    {"timestamp": "13:35", "noise": "Speech"},
    {"timestamp": "13:41", "noise": "Speech"},
    {"timestamp": "13:47", "noise": "Cough, Speech"}
  ]
}
```

```

{"timestamp": "13:52", "noise": "Speech"},
{"timestamp": "13:58", "noise": "Speech"},
{"timestamp": "14:03", "noise": "Speech, Animal"},
{"timestamp": "14:15", "noise": "Speech, Inside small room"},
{"timestamp": "14:21", "noise": "Speech, Laughter"},
{"timestamp": "14:27", "noise": "Speech"}
],
"gps_location": []
}

```

**Sensemaking agents – Sensemaking:** Between 13:23 and 14:29, the participant was predominantly sedentary. A brief walking bout occurred from approximately 13:24–13:26, with elevated step count ( $\approx 59$  steps/min), suggesting a short transition (e.g., walking to/from a meeting room). Outside this period, step counts remained very low, watch wear was continuous, no phone usage was detected, and ambient noise was consistently speech-dominated. Heart rate remained stable (77–83 bpm), consistent with light activity and conversation rather than exertion. GPS data were unavailable, so location context could not be confirmed.

**Presentation Agent – System Output:**

```

[
  {
    "start": "13:23",
    "end": "13:24",
    "posture": "sitting",
    "activity": "attending meeting, talking",
    "reasoning": "Very low step count, steady heart rate, and speech-related ambient noise indicate the participant was seated and engaged in conversation."
  },
  {
    "start": "13:24",
    "end": "13:26",
    "posture": "standing",
    "activity": "walking",
    "reasoning": "Step count increased to ~59 steps/min, indicating a brief walking transition."
  },
  {
    "start": "13:26",
    "end": "14:29",
    "posture": "sitting",
    "activity": "attending meeting, talking",
    "reasoning": "After the walking bout, low step counts resumed, speech remained present, and heart rate stayed stable, consistent with a seated meeting."
  }
]

```

## C APPENDIX C: AN EXAMPLE OF HOW GLOSS4HAR WORKS FOR TIMELINE RECONSTRUCTION TASK

**User Query:** On the day 2025/02/19, the subject pilot2 reported doing the following activities in chronological order:

activity, approximate\_duration

Wake up, 08:15

Morning routine (bathroom, walking around), 25 min

Riding bus, 20 min

Computer use, 5 min

Studying, 75 min

...

Go to sleep, 23:00

Can you tell me the list of postures and activities pilot2 did from **8am–9am** based on their passive sensing data and their reported timeline?

- Use the timeline to narrow down the activities they might be doing.
- There are no timestamps, so triangulate the list with the passive sensing data and use common sense to assign time periods.
- List activities for the entire one-hour period.
- They can do multiple activities at once. Suggest all the activities they might be doing.
- Make sure to include posture.
- Use dynamic time periods based on changes in activities and postures.
- Check the start and end time of each activity and posture, and make sure they are correct to the minute.
- If two annotations have the same posture and activity and are consecutive, merge them.
- Do not output labels with overlapping time.

### Action Plan Generation Agent – Action Plan:

- (1) Parse the query date and time window: [retracted date] 08:00:00 to 09:00:00.
- (2) Query the **location database** for GPS points; compute per-minute speed to distinguish stationary (home), walking, and vehicular movement.
- (3) Query the **step count database**; classify lying down (<5 steps/min), walking (5–30 steps/min), and seated transit (<2 steps/min).
- (4) Query the **phone usage database** to corroborate activity transitions.
- (5) Query the **ambient noise database** (every 5 min) to strengthen transit and indoor classifications.
- (6) Query the **watch wear database** to confirm data reliability across the interval.
- (7) Triangulate all signals to assign activity and posture labels per time segment.
- (8) Merge consecutive segments sharing the same activity and posture.
- (9) Output a non-overlapping JSON array of labeled segments.

**Memory:** (none retrieved for this query)

**Coding and Database Management Agents – Retrieved Sensor Data (08:00–09:00):**

```

{
  "step_count": [
    {"start_time": "08:25:38", "end_time": "08:39:05", "avg_steps_per_min": 89.67},
    {"start_time": "08:52:01", "end_time": "08:56:34", "avg_steps_per_min": 103.08}
  ],
  "phone_usage": [
    {"period": "08:16 to 08:19"},
    {"period": "less than a minute at 08:33"},
    {"period": "08:37 to 08:38"},
    {"period": "less than a minute at 08:42"},
    {"period": "less than a minute at 08:46"},
    {"period": "less than a minute at 08:49"},
    {"period": "less than a minute at 08:51"},
    {"period": "08:53 to 08:54"}
  ],
  "ambient_noise": [
    {"timestamp": "08:09", "noise": "Silence"},
    {"timestamp": "08:21", "noise": "Silence"},
    {"timestamp": "08:26", "noise": "Speech, Inside small room"},
    {"timestamp": "08:32", "noise": "Speech"},
    {"timestamp": "08:37", "noise": "Gunshot, gunfire"},
    {"timestamp": "08:42", "noise": "Speech, Fowl, Chicken, rooster"},
    {"timestamp": "08:48", "noise": "Inside, small room"},
    {"timestamp": "08:53", "noise": "Walk, footsteps"},
    {"timestamp": "08:59", "noise": "Insect, Fly, housefly"}
  ],
  "watch_non_wear": [
    {"start_time": "08:00:54", "end_time": "08:22:15"}
  ]
}

```

**Sensemaking agents – Sensemaking:** Between 08:00 and 09:00, step count shows two walking bouts: 08:25–08:39 ( $\approx 90$  steps/min) and 08:52–08:56 ( $\approx 103$  steps/min). GPS confirms eastward displacement across campus during both bouts. Phone usage was sporadic, with active screen usage at 08:16–08:19, 08:37–08:38, and 08:53–08:54, and brief touch events at 08:33, 08:42, 08:46, 08:49, and 08:51. Ambient noise transitions from silence (08:09, 08:21) to indoor speech (08:26, 08:32), outdoor noise (08:37, 08:42), and footsteps (08:53). The watch was not worn from 08:00:54 to 08:22:15; sensor data from 08:22 onward is reliable.

#### Presentation Agent – System Output:

```

[
  {"start": "08:00", "end": "08:15", "posture": "lying down",
   "activity": "sleeping",
   "reasoning": "Not yet awake per timeline; stationary GPS, no steps, silence."},
  {"start": "08:15", "end": "08:16", "posture": "standing",
   "activity": "getting ready",

```

```

    "reasoning": "User woke up; beginning morning routine."},
  {"start": "08:16", "end": "08:19", "posture": "standing",
   "activity": "phone using",
   "reasoning": "Phone screen active; no steps detected."},
  {"start": "08:19", "end": "08:25", "posture": "standing",
   "activity": "getting ready",
   "reasoning": "Continuation of morning routine before walking starts."},
  {"start": "08:25", "end": "08:40", "posture": "standing",
   "activity": "walking",
   "reasoning": "Sustained step count ~90 steps/min; GPS shows displacement."},
  {"start": "08:40", "end": "09:00", "posture": "sitting",
   "activity": "riding bus",
   "reasoning": "Timeline lists 20-min bus ride; rapid GPS displacement; near-zero steps."}
]

```

#### D APPENDIX D: SENSOR INPUT TO GLOSS4HAR

These are the raw data we hooked into the GLOSS4HAR system (under the database management agent). This is the system prompt for the agent. The  $\mu$ EMA data, is excluded in the reconstruction task for all conditions except for the  $\mu$ EMA condition.

This is the data description passed to **database management agent**:

```

{
  "location database": {
    "info": "Contains GPS location data (latitude, longitude, altitude) recorded via the
    phone.",
    "device": "Phone",
    "additional_instructions": "The location database can be used to detect activity related
    to the location, such as home, work, entertainment, etc. It can also detect speed to
    identify activity like riding train, bus, cycling, etc. Do all calculations in latitude
    and longitude values and call the address function only when you need to show the address
    to the user, as it is computationally expensive."
  },
  "uEMA database": {
    "info": "Contains user's self-report of in-the-moment activity on the watch. Some
    self-reports are voice-based, so responses are transcribed text.",
    "device": "Watch",
    "additional_instructions": "Some transcribed responses might not be accurate (e.g.,
    'sitting' transcribed as 'setting'). Use the uEMA database to retrieve what activity the
    participant did at that moment, not the start time. To get exact start and end times,
    first use other databases to find a consistent time period, then match with uEMA for the
    activity label."
  },
  "heart rate database": {
    "info": "Contains heart rate data (anytime it changes) recorded from the Pixel
    smartwatch.",
    "device": "Watch",

```

```

    "additional_instructions": "Used to understand the user's physical activity level.
    Combined with step count, it can detect moderate to vigorous activities that do not
    involve a lot of steps."
  },
  "phone usage database": {
    "info": "Contains phone usage data. True means the phone is being used; false means it is
    not being used at the time collected.",
    "device": "Phone",
    "additional_instructions": "Used to detect 'phone using' activity."
  },
  "ambient noise database": {
    "info": "Contains ambient noise data recorded from the Pixel smartwatch every 5 minutes.",
    "device": "Watch",
    "additional_instructions": "Noises are detected using the Google YAMNet model, which
    classifies noise into 521 classes. Some predictions may be inaccurate or unrelated to the
    user's activity."
  },
  "step count database": {
    "info": "Contains step count data recorded from the Pixel smartwatch every minute.",
    "device": "Watch",
    "additional_instructions": "Used to detect activities such as walking and running.
    Combined with heart rate or location, it can detect activities that do not involve
    walking or running, such as cycling or riding in a car."
  },
  "skin temperature database": {
    "info": "Contains skin temperature data recorded from the Pixel smartwatch every minute.",
    "device": "Watch",
    "additional_instructions": "Can be used to detect activities that involve changes in skin
    temperature, such as exercise or sleep."
  },
  "watch wear database": {
    "info": "Contains data on whether the watch is worn or not. True means worn; false means
    not worn at the time collected.",
    "device": "Watch",
    "additional_instructions": "Used to detect activities such as wearing the watch, taking
    it off, or putting it on."
  }
}

```

## E APPENDIX E: ABLATION OF SYSTEM COMPONENTS

We conducted ablation experiments to study the impact of removing GLOSS4HAR previous predictions memory and closed-ended activity labels.

We compared three conditions: 1) full GLOSS4HAR system with all components, 2) full GLOSS4HAR system, but without memory of past predictions (so each run is independent of each other), and 3) full GLOSS4HAR system, but removing the list of postures and activities (Table 2) from the presentation agent.

We report the results of the ablation study in Table 7 and Table 8. The results show that both the memory component and the closed-end label constraint substantially impact performance in the annotation correction

task. In the reconstruction task, the closed-end label constraint consistently improves performance across all conditions. In contrast, the memory component has little effect in the list-of-activities conditions. We hypothesize that this is because, in these conditions, GLOSS4HAR has access to a full-day timeline at each query (i.e., each one-hour window includes the participant’s entire self-reported activity list for that day), providing sufficient temporal context. This contrasts with the correction task, where the model only observes a limited snippet of the participant’s self-reports (one annotation at a time).

		Correction Task			Reconstruction Task											
					No self-reports			$\mu$ EMA			List (approx)			List (no time)		
		EM	F1	Jl	EM	F1	Jl	EM	F1	Jl	EM	F1	Jl	EM	F1	Jl
GLOSS4HAR	gpt-oss	0.73	0.92	0.87	0.38	0.74	0.67	0.38	0.77	0.68	0.51	0.81	0.74	0.45	0.78	0.71
w/o memory		0.62	0.88	0.82	0.25	0.72	0.61	0.13	0.68	0.56	0.56	0.84	0.77	0.45	0.80	0.72
w/o closed-end labels		0.44	0.82	0.72	0.18	0.70	0.58	0.13	0.68	0.56	0.35	0.76	0.67	0.28	0.73	0.62
GLOSS4HAR	gpt-4o	0.76	0.94	0.89	0.34	0.76	0.66	0.44	0.79	0.71	0.65	0.87	0.81	0.47	0.81	0.73
w/o memory		0.62	0.87	0.81	0.27	0.73	0.62	0.36	0.77	0.67	0.71	0.90	0.85	0.45	0.80	0.71
w/o closed-end labels		0.48	0.86	0.74	0.37	0.70	0.65	0.14	0.68	0.56	0.33	0.75	0.64	0.40	0.75	0.67

Table 7. Performance comparison on correction and reconstruction tasks, against researcher-corrected ground truth.

		No self-reports			$\mu$ EMA			List (approx)			List (no time)		
		EM	F1	Jl	EM	F1	Jl	EM	F1	Jl	EM	F1	Jl
GLOSS4HAR	gpt-oss	0.33	0.73	0.65	0.29	0.76	0.65	0.46	0.80	0.71	0.40	0.78	0.69
w/o memory		0.25	0.72	0.61	0.12	0.68	0.56	0.56	0.84	0.77	0.45	0.80	0.72
w/o closed-end labels		0.18	0.70	0.58	0.13	0.68	0.56	0.35	0.76	0.67	0.28	0.73	0.62
GLOSS4HAR	gpt-4o	0.26	0.75	0.62	0.33	0.74	0.65	0.61	0.86	0.80	0.45	0.81	0.71
w/o memory		0.27	0.73	0.62	0.37	0.77	0.68	0.71	0.90	0.85	0.45	0.80	0.71
w/o closed-end labels		0.37	0.70	0.65	0.14	0.68	0.56	0.33	0.75	0.64	0.40	0.75	0.67

Table 8. Performance comparison on reconstruction tasks, against participants’ original annotations.

## F APPENDIX F: MODEL COST AND LATENCY

Model cost and latency is important for assessing reproducibility and real-world deployment feasibility. We report model cost and latency in Table 9, including input and output token usage, wall-clock execution time, for both the gpt-4o and gpt-oss models. At the time of the experiments, gpt-4o is priced at \$2.50 per million input tokens and \$10.00 per million output tokens. For gpt-oss, the model is executed on a local GPU, and thus the monetary cost is fixed and independent of token count.

	Correction Task			Reconstruction Task		
	Input tokens	Output tokens	Latency (s)	Input tokens	Output tokens	Latency (s)
gpt-oss	1237.3 ± 32.0	10901.1 ± 5615.6	10.2 ± 3.4	1397.0 ± 49.3	14128.4 ± 9635.0	12.3 ± 3.4
gpt-4o	1241.7 ± 25.4	11588.4 ± 4727.5	24.0 ± 11.3	1430.3 ± 34.4	10029.3 ± 4375.3	22.8 ± 11.3

Table 9. Token usage and latency for correction and reconstruction tasks. Token count and latency is per query, for the entire system (all agents). For the correction task, the number of query per day is equal to the number of annotations. For the reconstruction task, the number of query per day is 24 (one query per hour).

## G APPENDIX G: THIGH-SENSOR HAR MODEL

The thigh-sensor predictions are derived from accelerometer data from the Actigraph sensor. We ran the accelerometer data through a HAR model that have been pre-trained on the PAAWS dataset [77]. The PAAWS dataset was collected from 252 participants engaged in 33 laboratory activities around a university campus. We used data from 38 participants in the PAAWS dataset to train the model. Since participants in the ACAI study could choose to wear the sensor on either their left or right thigh, we used the tri-axial acceleration data sampled at 80 Hz from an Actigraph GT9X-Link (the same model as the sensor used in ACAI study with similar placement) placed on the participants' left and right thighs in the PAAWS dataset. For training and evaluation, we segmented the raw acceleration data into 60 s (4,800 samples), non-overlapping windows and extracted 11 features from the accelerometer signal (66 features used in the feature vector). We implemented our RF model using sklearn v1.0.2, 1,000 trees, and otherwise default settings. Table 10 shows the results of validating the random forest model on the PAAWS dataset.

Table 10. The results of leave-one-participant-out cross validation of the random forest (RF) model on the PAAWS dataset. Results are reported by thigh side (Right and Left) and by average F1.

Side	Lying_Down	Running	Sitting	Standing	Walking	Overall
Right	0.86 ± 0.05	0.56 ± 0.31	0.91 ± 0.03	0.79 ± 0.06	0.91 ± 0.02	0.87 ± 0.02
Left	0.86 ± 0.04	0.54 ± 0.32	0.91 ± 0.03	0.79 ± 0.06	0.91 ± 0.02	0.87 ± 0.03
Avg	0.86 ± 0.04	0.56 ± 0.32	0.91 ± 0.03	0.79 ± 0.06	0.91 ± 0.02	0.87 ± 0.03

## H APPENDIX H: AN EXAMPLE OF HOW THE RAG BASELINE WORKS FOR ANNOTATION CORRECTION

This appendix presents how the RAG baseline works for the same input query used in Appendix B, enabling direct comparison with GLOSS4HAR. We used the same RAG implementation as the baseline used in the original GLOSS paper [19].

**User Query:** Subject with id pilot7 annotated sitting, attending meeting during the entire period from 13:23–14:29 on 2025/02/27.

Based on passive sensing data, provide more accurate annotations:

- Check 10 minutes before and after the interval to refine start/end times.
- Correct posture and activities if needed.
- If the participant did multiple things, include all relevant activities.
- Merge consecutive identical labels.

- Do not output overlapping annotations.

**Query Parsing – Extracting retrieval parameters:**

Task: Identify the relevant time interval and sensor databases needed to answer the query.

Input:

- User query describing annotation to be corrected
- List of available sensor databases (e.g., step count, heart rate, phone usage, ambient noise)

Output:

- User ID: pilot7
- Start time: 2025-02-27 13:13:00
- End time: 2025-02-27 14:39:00
- Databases: step count, heart rate, ambient noise, phone usage

**Note:** This step retrieves raw sensor data from the databases according to the params and pass it and pass it to the next step to get a natural language narrative.

**Retrieval – Retrieved sensor data converted to natural language:**

Between 13:13 and 14:39, the participant showed mostly low step counts, with a short period of increased movement around 13:24–13:26 suggesting brief walking. Heart rate remained stable between 77–83 bpm. Ambient noise was consistently classified as speech, with occasional laughter and indoor sounds. No phone usage was detected during this period.

The participant appears to have been mostly stationary with intermittent conversational activity in an indoor environment.

**RAG Prompt – Question answering with retrieved context:**

You are an assistant for question-answering tasks for data from various sensors. Use the following context to answer the question. If you do not know the answer, say so. You may make reasonable estimates if exact information is unavailable.

**Question:** Refine the participant’s annotation from 13:23–14:29.

**Context:** [Retrieved sensor narrative]

**Answer:**

**Model Output:**

```
[
  {
    "start": "13:23",
    "end": "13:29",
    "posture": "sitting",
    "activity": "attending meeting, talking"
  }
]
```