

SensorChat: Answering Qualitative and Quantitative Questions during Long-Term Multimodal Sensor Interactions

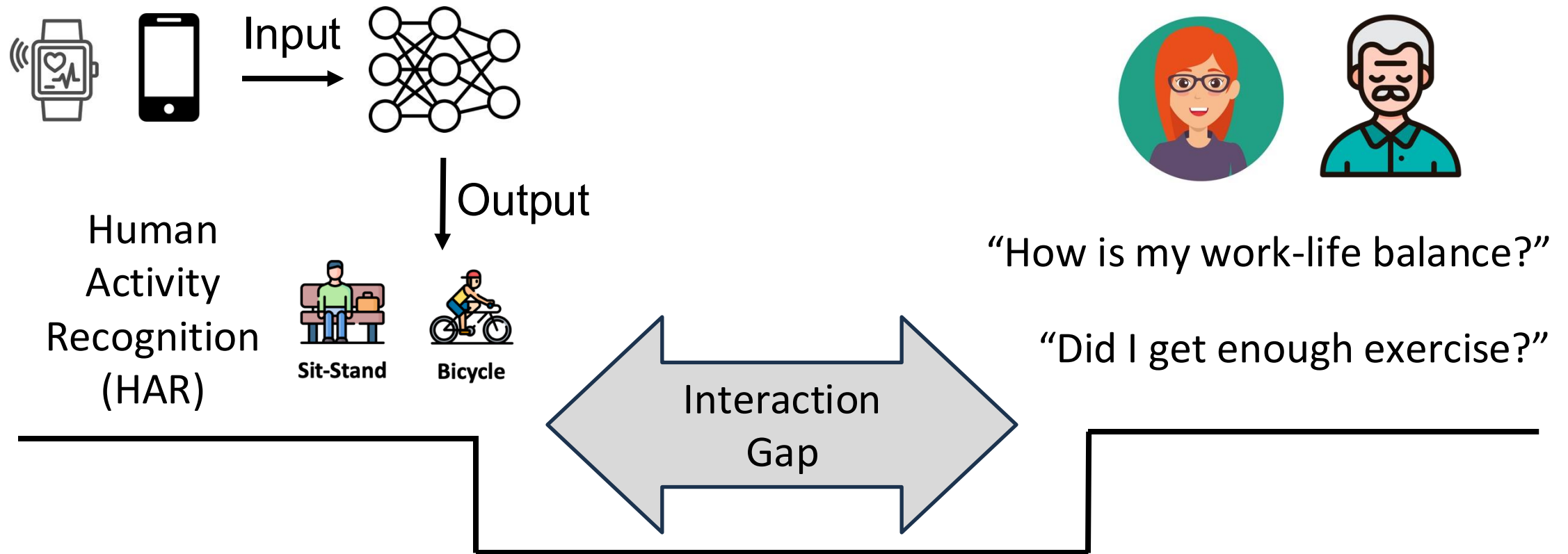
Xiaofan Yu¹, Lanxiang Hu¹, Benjamin Reichman², Dylan Chu¹, Rushil Chandrupatla¹, Xiyuan Zhang¹, Larry Heck², Tajana Rosing¹

¹ University of California San Diego

² Georgia Institute of Technology

x1yu@ucsd.edu

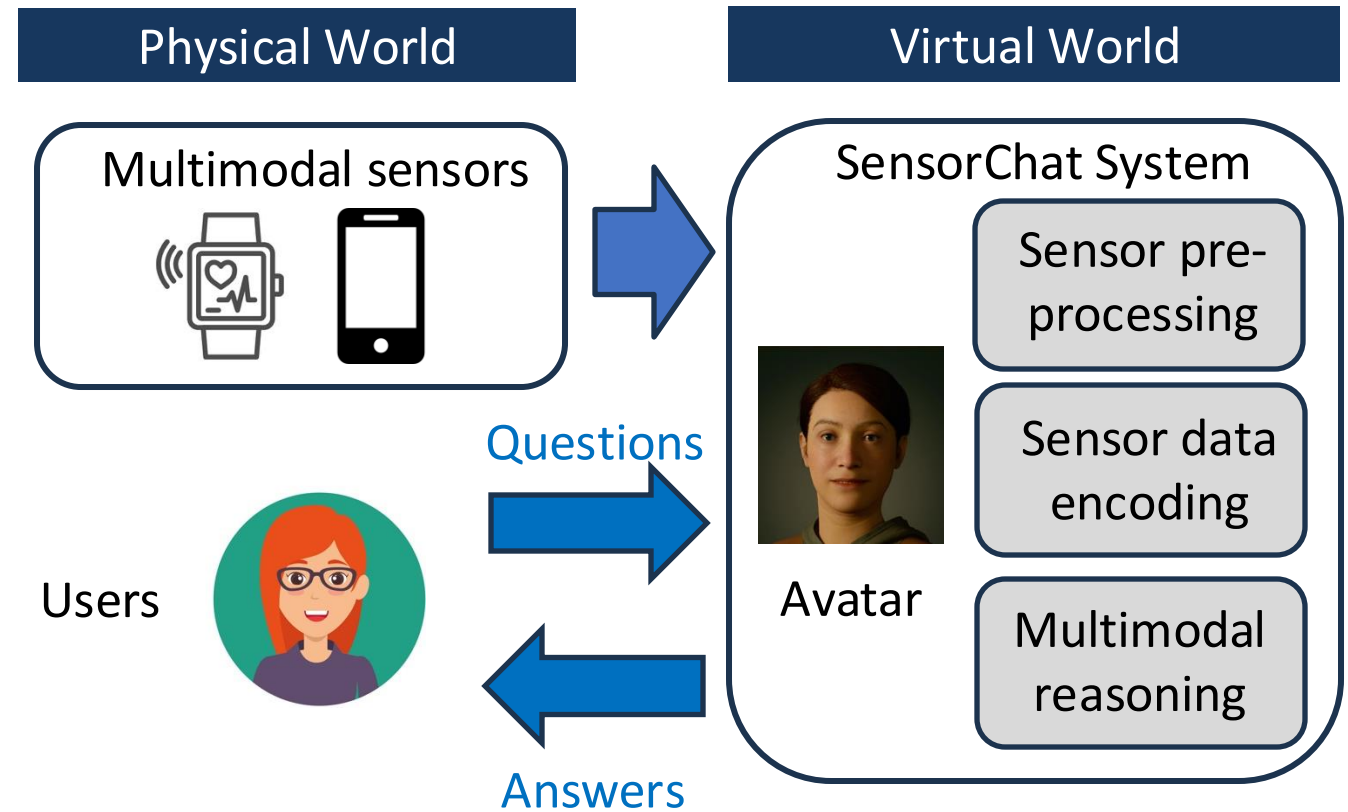
Motivation



How to make sensor data more accessible and useful to people?

Natural Language Interaction with Sensors

- Question-answering interactions with sensors
 - **Input:** multimodal sensor data and arbitrary questions from users
 - **Output:** answers to users
- Free-form chatting with sensors becomes possible with Large Language Models (LLMs)



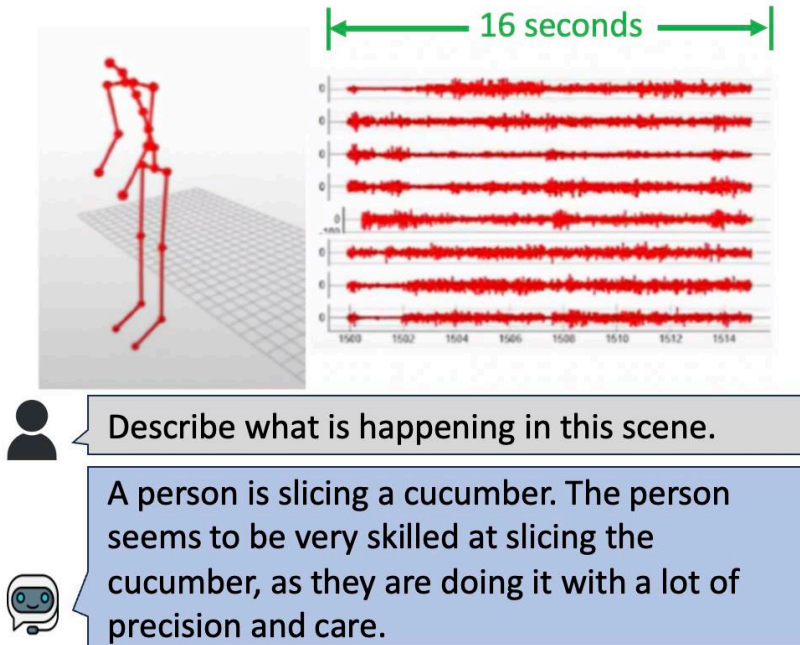
Problem Definition

- We focus on practical Q&As that involve *long-duration*, *high-frequency* sensor data, and *qualitative* and *quantitative* questions

Table 1. Key concepts that defines the problem scope of this paper. These concepts are motivated by practical user needs.

Concept	Meaning
Long-duration data	Sensor data spans more than a full day of user activity
High-frequency data	Collected from high-frequency sensors (e.g., IMU raw signals), as opposed to sparse data like daily step counts or statistics (e.g., mean value)
Quantitative questions	Require precise and objective answers based on sensor data, such as true/false, location, activity, count, or time duration
Qualitative questions	Require high-level reasoning to infer subjective insights, such as work-life balance and social interactions, possibly involving external knowledge

State of the Art



DeepSQA [IoTDI'23], OneLLM [CVPR'24],
Sensor2Text [IMWUT'24]

☹️ Limited sensor time range

Sensor data recordings: Temp: 37°C, HR: 85 bpm, Resp. rate: 16 bpm. No fever (>38°C), high pulse (>100 bpm), or rapid breathing (>22 bpm).

I've been coughing for two days and have yellow phlegm in my throat.

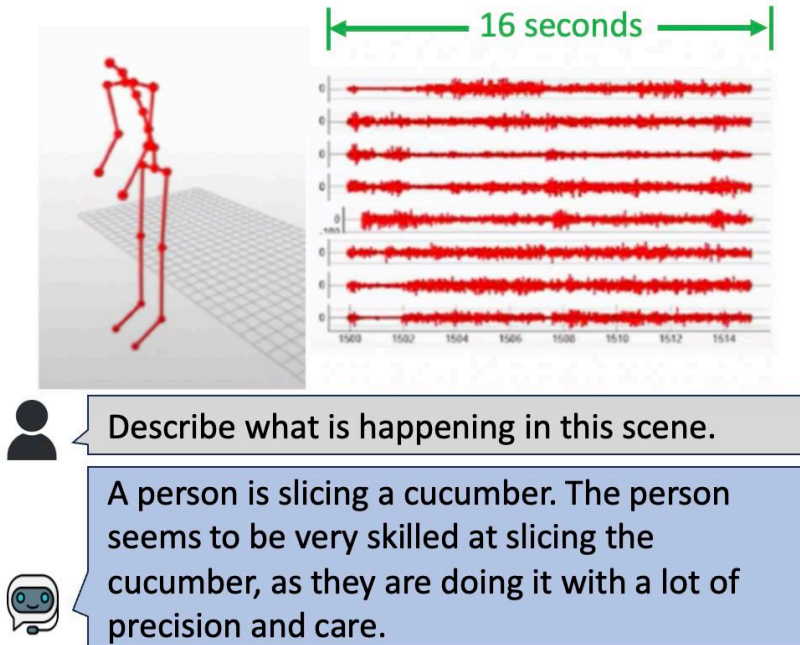
.....

Based on the information provided and the absence of concern for a specific pathogen that would change management, a clinical diagnosis of acute bronchitis can be made.

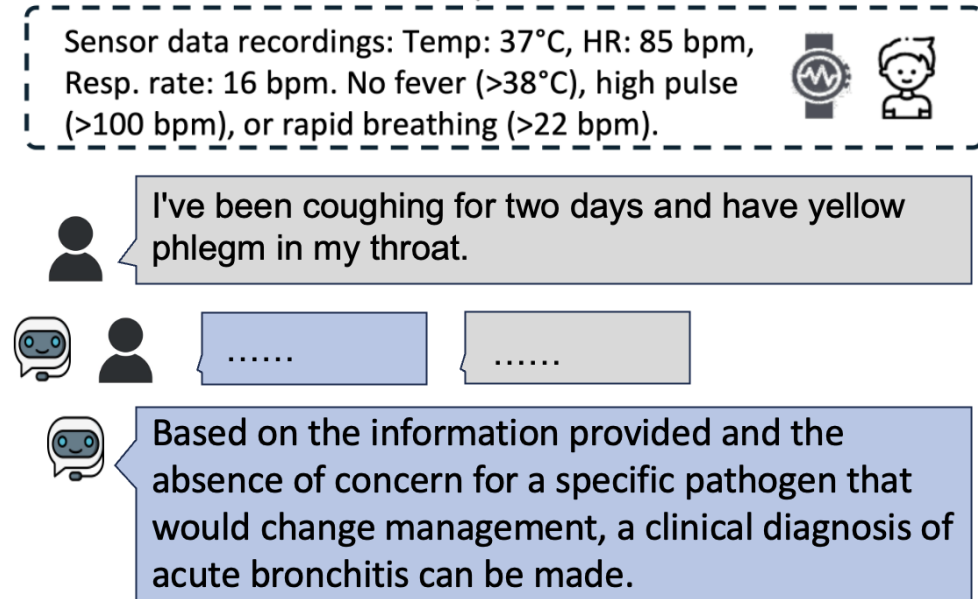
Health-LLM [PMLR'24],
DrHouse [IMWUT'24]

☹️ Low-dimensional sensor data

State of the Art



DeepSQA [IoTDI'23], OneLLM [CVPR'24],
Sensor2Text [IMWUT'24]





Health-LLM [PMLR'24],
DrHouse [IMWUT'24]

No existing QA benchmark for sensors has included **long-duration, high-dimensional** sensor data!

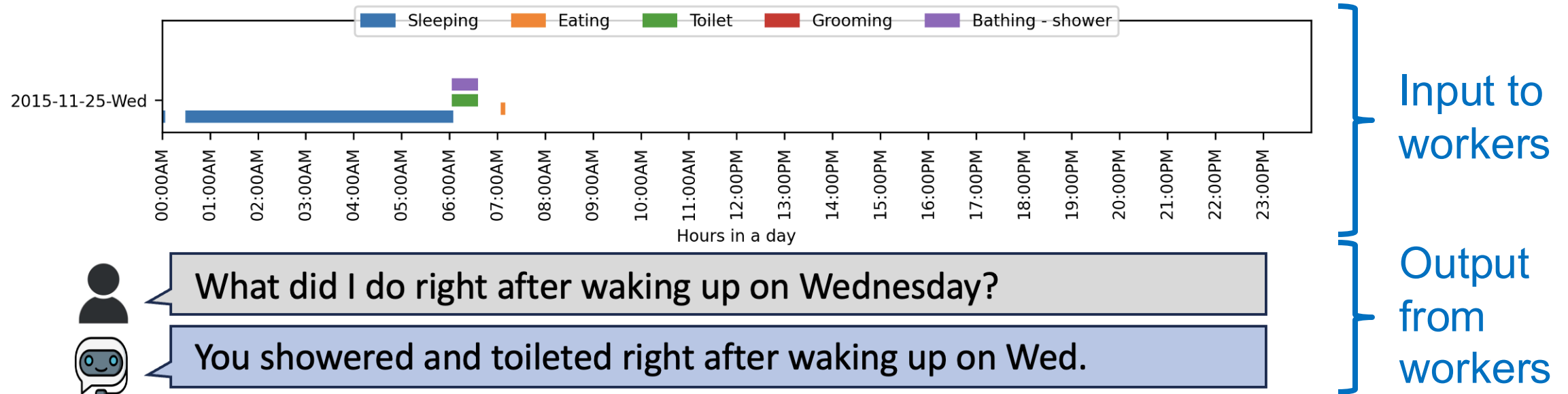
Our Contribution: SensorQA Dataset [SenSys'25]

- Introducing SensorQA, a human-created QA dataset for long-duration multimodal sensors, aimed at real-world scenarios

Goals	SensorQA Design
 <ul style="list-style-type: none"> • Naturally collected sensor data with long time span 	<ul style="list-style-type: none"> • Sensor data from ExtraSensory [IMWUT'17] <ul style="list-style-type: none"> • IMUs on phone & watch, audio (MFCC), GPS, compass, phone status, etc • 60 users, 51 activity labels, 2-10 days
 <ul style="list-style-type: none"> • Diverse questions and answers that align with human interests 	<ul style="list-style-type: none"> • Crowdsourcing Q&A pairs using Amazon Mechanical Turk¹ <ul style="list-style-type: none"> • Multi-time scale activity graph • 14 label subsets on different life aspects

Our Contribution: SensorQA Dataset (Cont.)

- Collected 5,648 Q&A pair generated by the AMT human workers



- Diverse Q&As include time queries, day queries, counting, activity queries
- Correctly answering the questions may require multi-step multimodal reasoning and quantitative analysis

How does SOTA Perform on SensorQA?

- **Answer accuracy:** matching key phrases in the generated vs. true answers

E.g.

Shortened True Answer: **Wednesday.**

Correct Answer: You had the least meals on **Wednesday.**

Wrong Answer: You had the longest eating session on **Thursday.**

Modality	Method	Backbone	Answer Accuracy
Text	LoRA finetuning	LLaMA2-7B	0.27
Image+Text	GPT4o	-	0.20
Sensor+Text	IMU2CLIP + GPT-4 [EMNLP'23]	GPT-4	0.13
Sensor+Text	DeepSQA [IoTDI'21]	CNN+LSTM	0.27
Sensor+Text	OneLLM [CVPR'24]	LLaMA2-7B	0.05

How does SOTA Perform on SensorQA?

- **Answer accuracy:** matching key phrases in the generated vs. true answers

E.g.

Shortened True Answer: **Wednesday**.

Correct Answer: You had the least meals on **Wednesday**.

Wrong Answer: You had the longest eating session on **Thursday**.

Modality	Method	Backbone	Answer Accuracy
Text	LoRA finetuning	LLaMA2-7B	0.27
Image+Text	GPT4o	-	0.20
Sensor+Text	IMU2CLIP + GPT-4 [EMNLP'23]	GPT-4	0.13
Sensor+Text	DeepSQA [IoTDI'21]	CNN+LSTM	0.27
Sensor+Text	OneLLM [CVPR'24]	LLaMA2-7B	0.05

How does SOTA Perform on SensorQA?

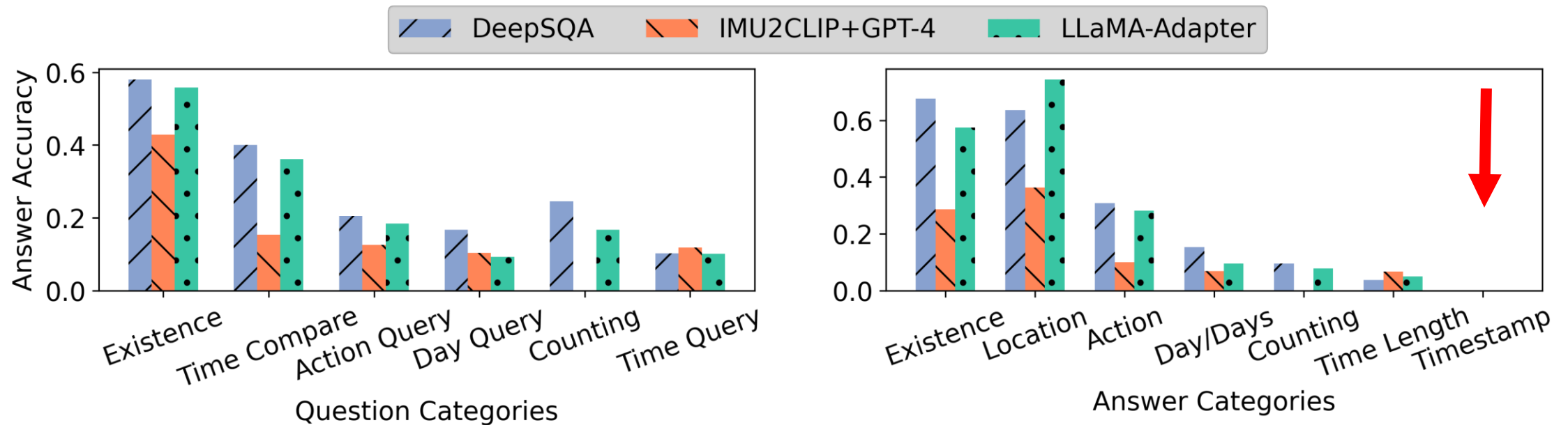
- **Answer accuracy:** matching key phrases in the generated vs. true answers

Modality	Method	Backbone	Answer Accuracy
Text	LoRA finetuning	LLaMA2-7B	0.27
Image+Text	GPT4o	-	0.20
Sensor+Text	IMU2CLIP + GPT-4 [EMNLP'23]	GPT-4	0.13
Sensor+Text	DeepSQA [IoTDI'21]	CNN+LSTM	0.27
Sensor+Text	OneLLM [CVPR'24]	LLaMA2-7B	0.05

Lesson 1: Ineffective multimodal fusion leads to poor answer accuracy

How does SOTA Perform on SensorQA? (Cont.)

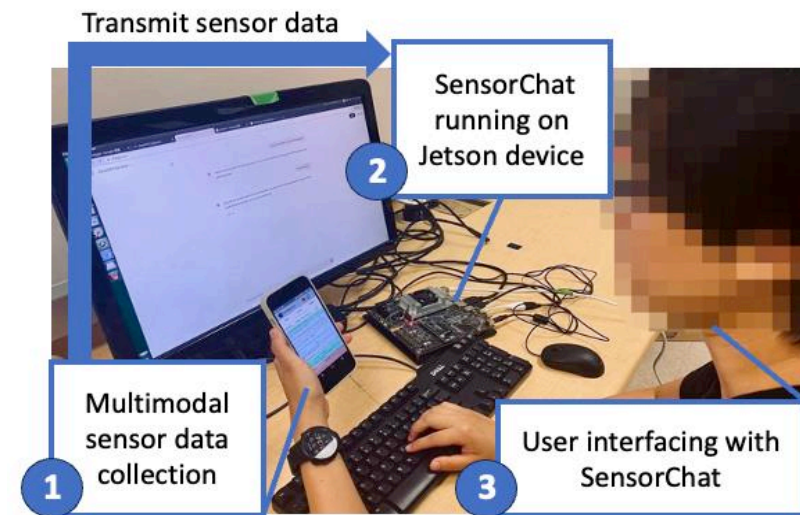
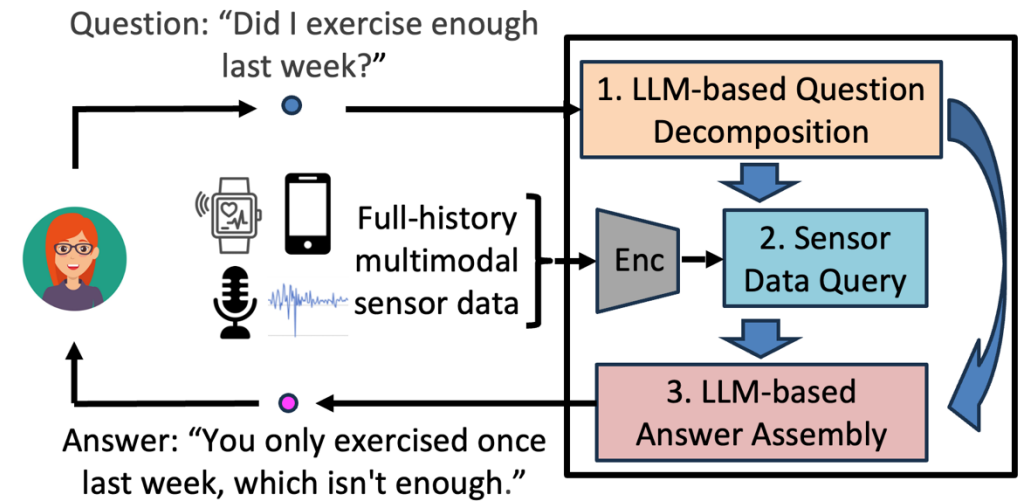
- Profile the accuracy **per question and answer category**



Lesson 2: SOTA methods struggle with accurate quantitative answers

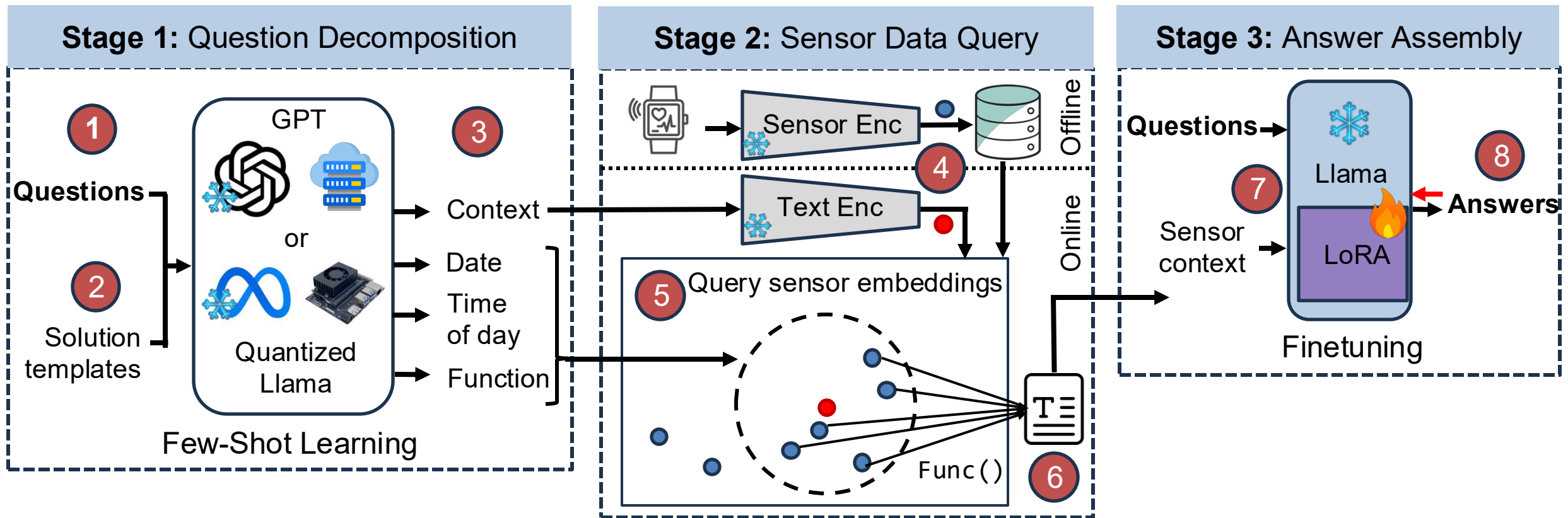
My Contribution: SensorChat [Yu et al., arXiv'25]

- Introducing SensorChat: a novel three-stage QA system for **long-duration, high-dimensional** multimodal sensor data
- Handling both quantitative and qualitative questions, achieving up to 93% higher accuracy than SOTA
- SensorChat system deployment
 - Real-time interactions on a cloud server
 - Running locally on NVIDIA Jetson Orin



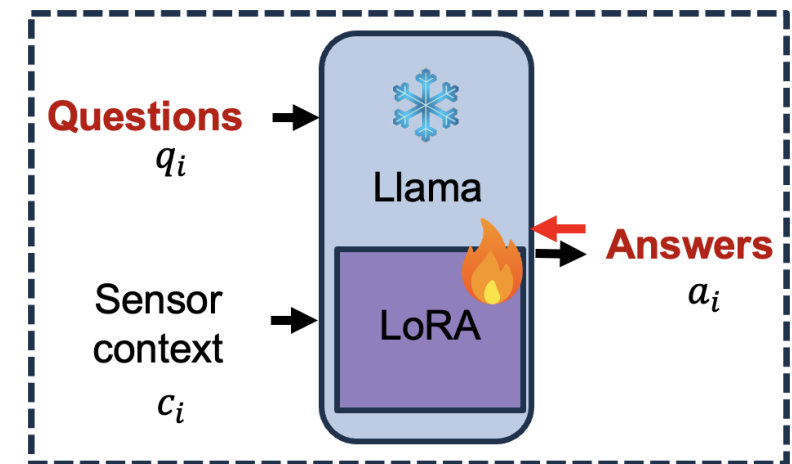
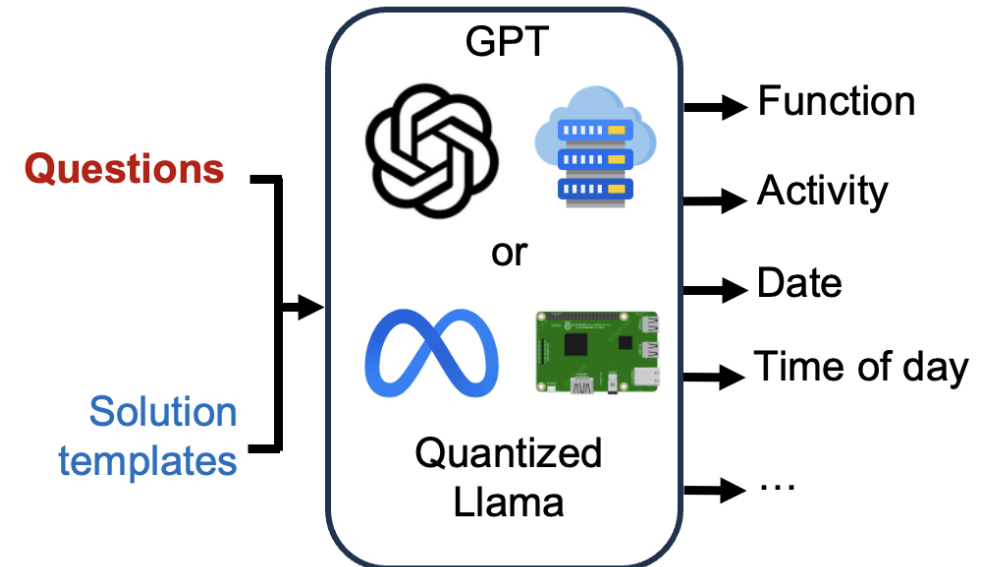
Overview of SensorChat

- **Goal:** Accurately answering based on long-duration sensor data
- **My solution in SensorChat:** LLM understanding + explicit sensor data query



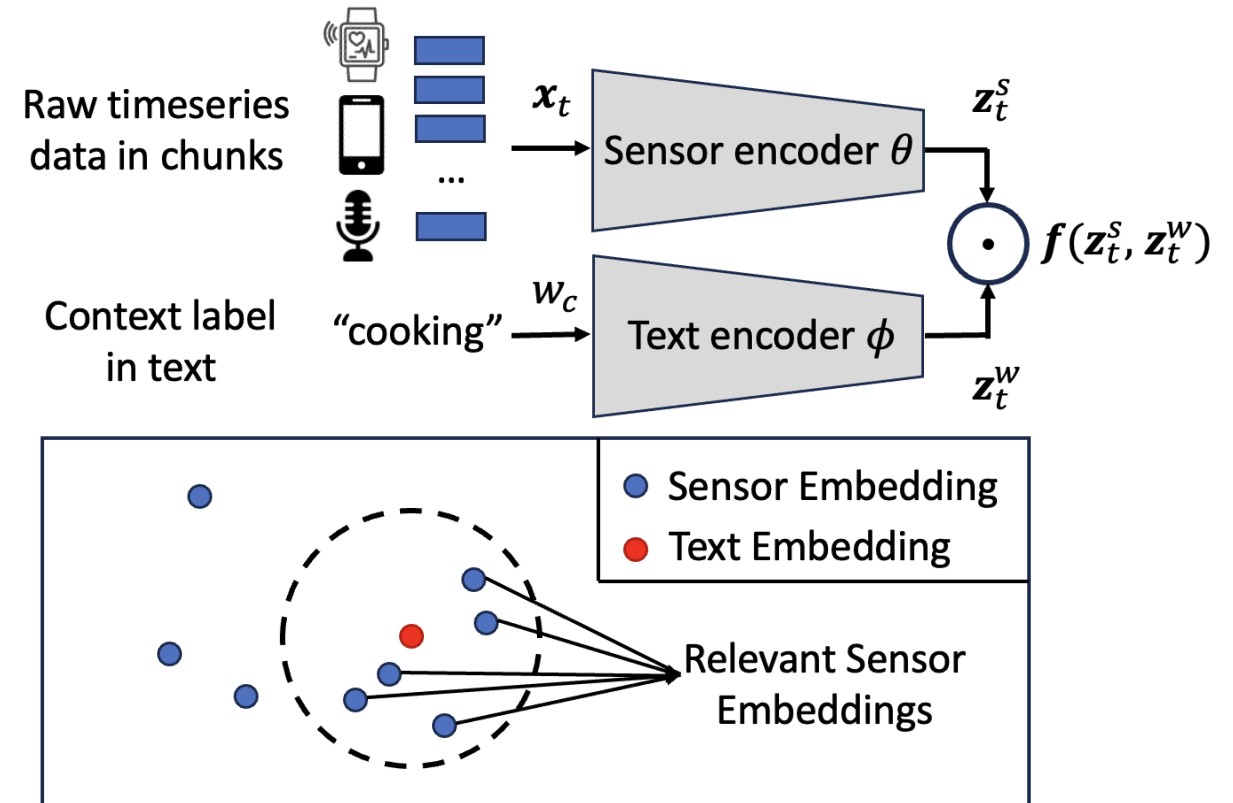
Question Decomposition & Answer Assembly

- Both stages are based on LLMs
- Question decomposition
 - Decompose questions into query functions
 - **In-context learning:** use solution templates to assist accurate decomposition per question type
 - **Chain-of-Thought (CoT) reasoning:** explicit require step-by-step reasoning
- Answer assembly
 - Produce answers based on question and query results
 - The finetuned and quantized LLaMA model runs locally on the edge device



Explicit Sensor Data Query Stage

- **Goal:** Accurately querying from long-duration sensor data
- Contrastive sensor-text pretraining
 - A novel loss function
- Effective query search in the database
 - A set of query functions



Example query function:

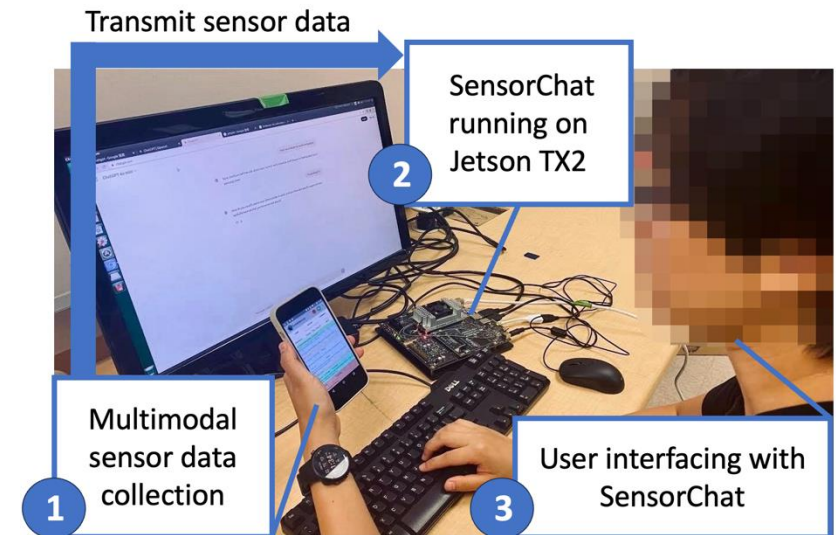
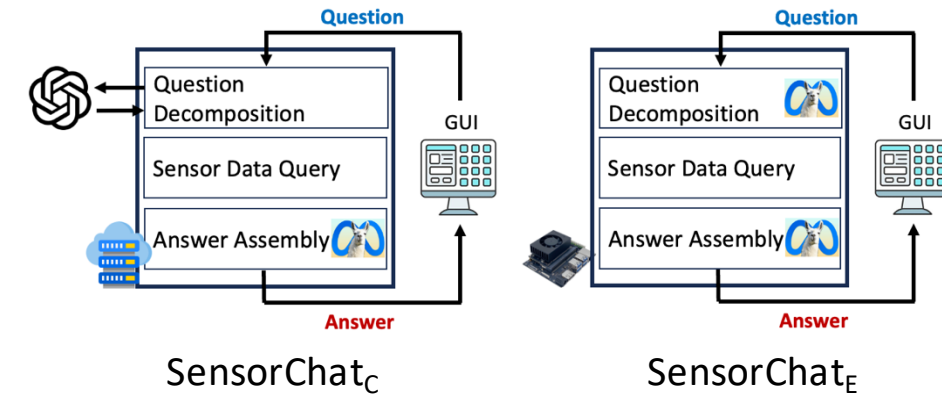
Question: “*How long* did I spend cooking last Sunday morning?”

Query output: “You spend γ minutes cooking on Sunday morning.”

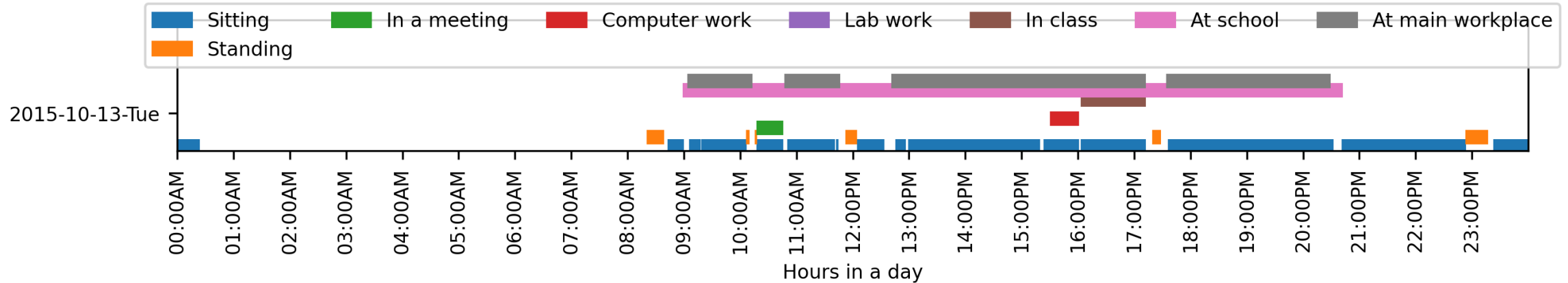
$$\gamma = \sum_{t \in T_{\text{SundayMorning}}} \left[f(\theta(\mathbf{x}_t), \phi(\text{"cooking"})) > h \right]$$

Experimental Setup

- Hardware Platform: A100, NVIDIA Jetson Orin
- SOTA Baselines:
 - **Pretrained methods:** GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, GPT-4o
 - **Trained or Finetuned methods** using LoRA [ICLR'22]
 - **Language-only methods:** T5 [JMLR'21], LLaMA [arXiv'23]
 - **Vision-based methods:** LLaMA-Adapter [ICLR'24], Llava-1.5 [arXiv'23]
 - **Multimodal methods:** DeepSQA [IoTDI'21], IMU2CLIP+GPT-4 [EMNLP'23], OneLLM [CVPR'24]
- Metrics
 - Full answer quality: Rouge scores
 - Answer accuracy: exact match scores on key answer phrases
 - Efficiency: memory requirement, generating latency per answer



SensorChat vs. SOTA QA Methods



Question: How much more time did I spend at school vs the main workplace?
True Answer: You spent **2 more hours, 15 min** at school vs the main workplace

- SensorChat produces natural and accurate answers to challenging questions



GPT4o: Based on the provided chart, you spent approximately 6.5 hours at school and approximately 5 hours at the main workplace. To fine the difference in time spent, you spent about **1.5 hours** more at school compared to the main workplace
LLaMA text only: You spent about **15 more minutes** at school vs the main workplace
LLaMA-Adapter: You spent **1 hour and 15 minutes** at school today
DeepSQA: **4 hours**

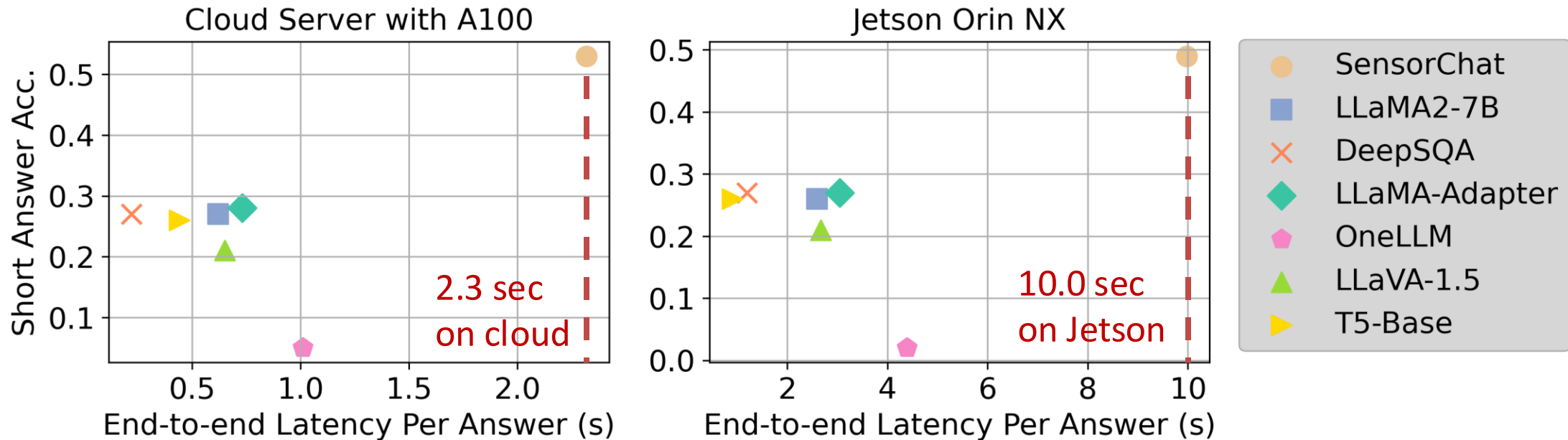
Understand Key Designs in SensorChat

Method	Answer Accuracy
Full SensorChat	0.54
w/o Question Decomposition	0.36
w/o Sensor Data Query	0.27
w/o Answer Assembly	0.0

- Good performance is resulted from the collaborative effort of all three stages
- The three-stage design offers effective fusion between the long-duration, high-dimensional sensor data and the textual questions

Inference Latency

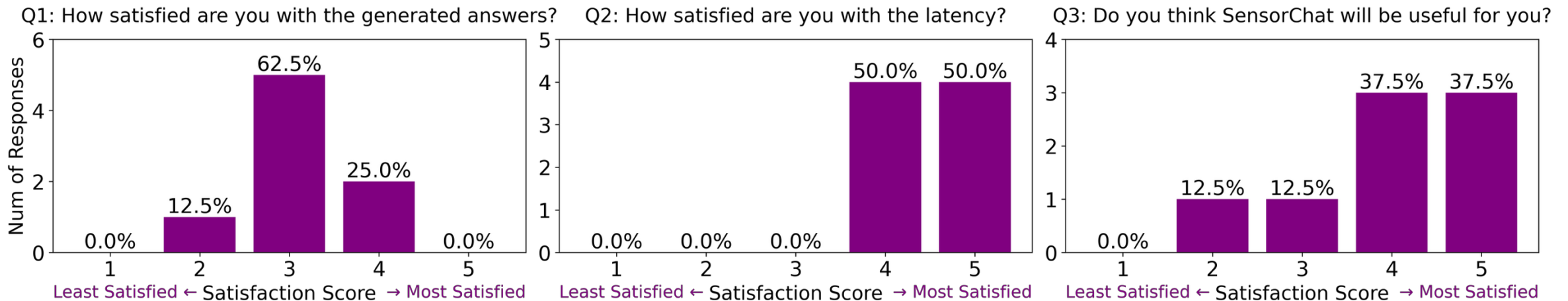
- SensorChat_C: Full-precision LLMs running on A100
- SensorChat_E: Quantized (4-bit) LLMs running on NVIDIA Jetson Orin



- SensorChat outperforms other baselines in accuracy while maintaining reasonable latency per answer on both cloud and edge devices

Real-World User Study

- We recruited eight volunteers to use SensorChat_c after carrying a smartphone for 1-3 days
- We collected user satisfactory scores from three perspectives



- SensorChat received an average score of 3.12 for answer content, 4.50 for latency, and 4.00 for practical utility

Summary and Impact

- Natural language interaction is key to make sensor data more accessible and useful to human users
- Prior works have limited sensor data time range and complexity
- We introduce SensorChat, a novel three-stage end-to-end system for real-time natural language interactions between humans and multimodal sensors
- Extensive experiments demonstrate SensorChat's practicality and efficiency on edge platforms
- Dataset is available at: <https://github.com/benjamin-reichman/SensorQA>

