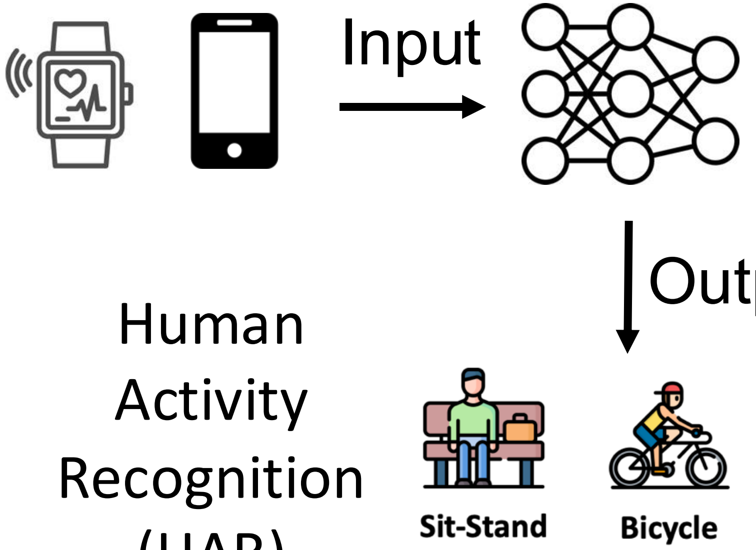
# SensorQA: A Question Answering Benchmark for Daily-Life Monitoring

Benjamin Reichman*[1], Xiaofan Yu*[2], Lanxiang Hu[2], Jack Truxal[1], Atishay Jain[1], Rushil Chandrupatla[2], Tajana Rosing[2], Larry Heck[1]

[1] Georgia Institute of Technology
[2] University of California San Diego
* Equal Contributions

SenSys 2025

# Motivation




Human Activity Recognition (HAR)

"How is my work-life balance?"

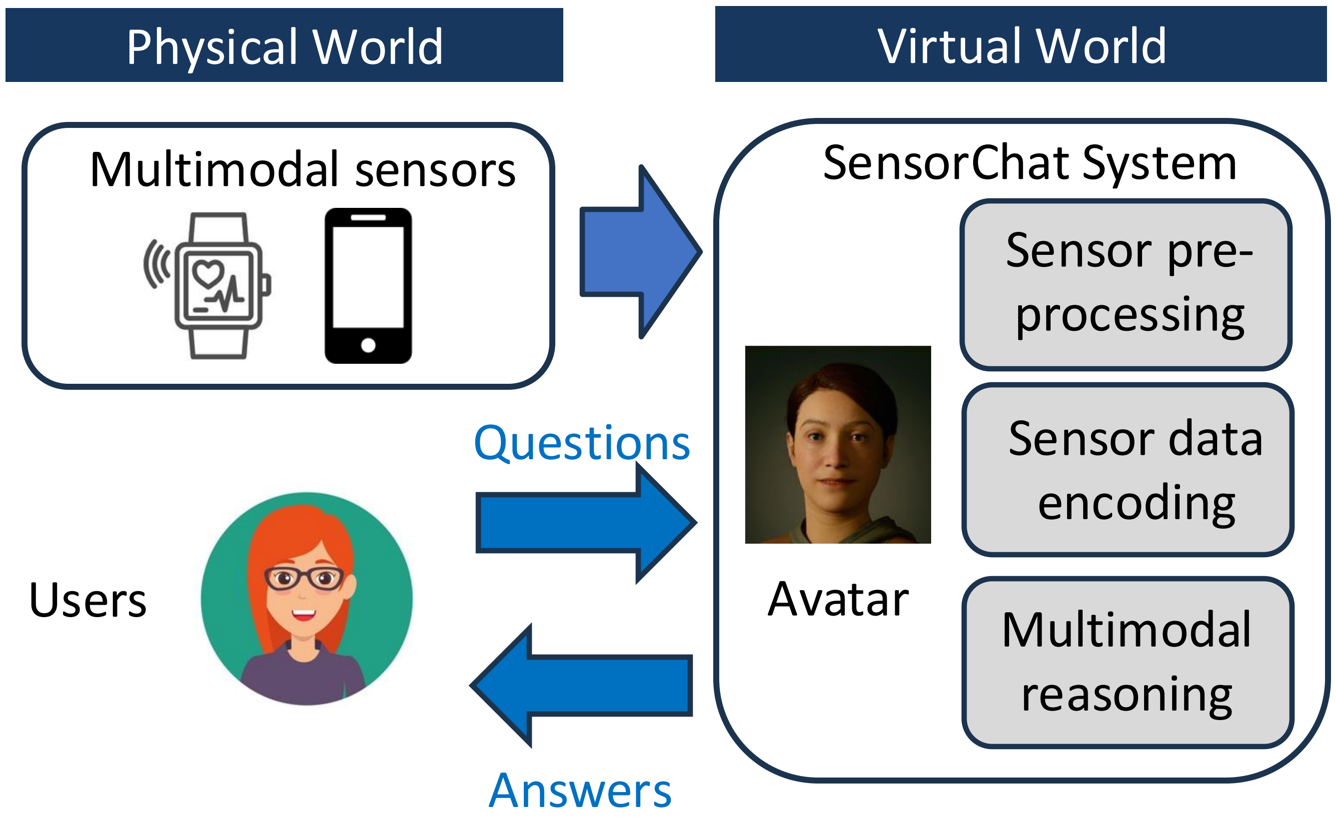
"Did I get enough exercise?"

Interaction Gap

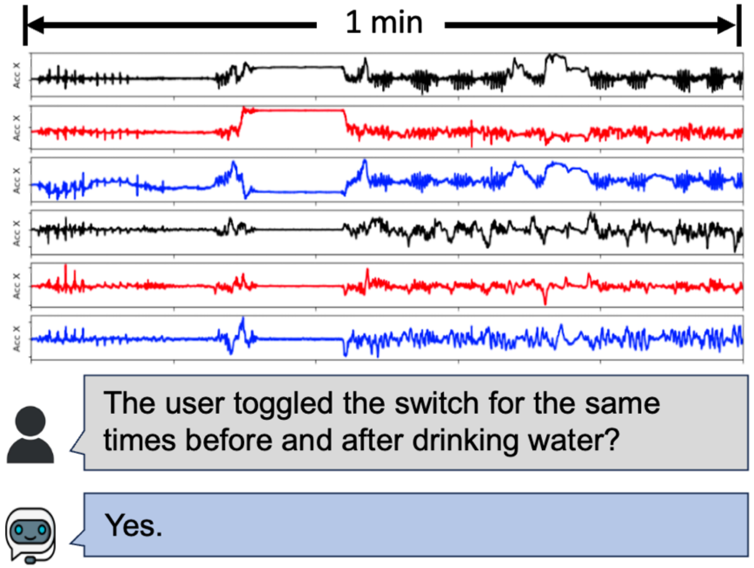
How to make sensor data more accessible and useful to people?

# Natural Language Interaction with Sensors

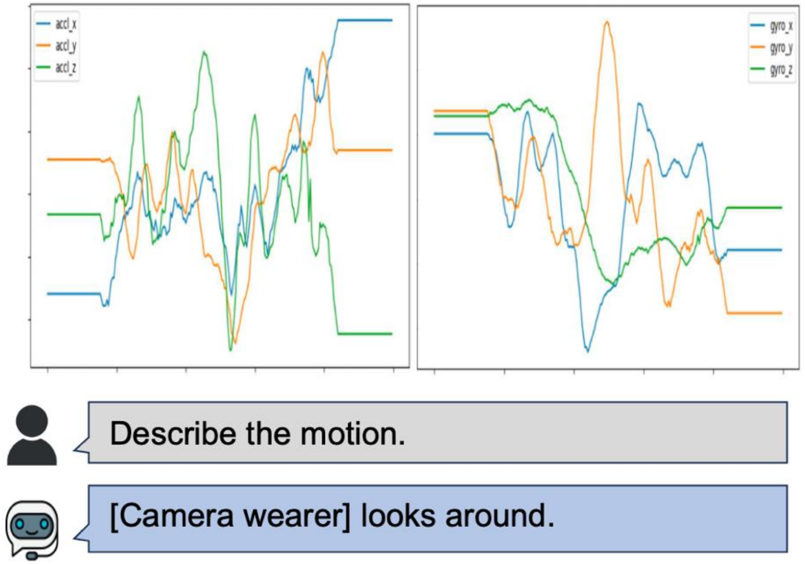
- Question-answering interactions with sensors
  - **Input:** multimodal sensor data and arbitrary questions from users
  - **Output:** answers to users
- Free-form chatting with sensors becomes possible with Large Language Models (LLMs)



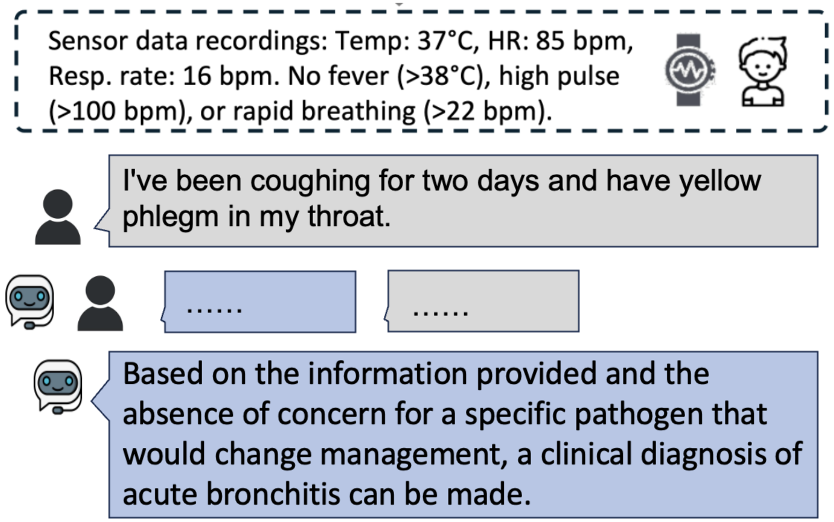
# State of the Art




DeepSQA [IoTDI’21]




AnyMAL [arXiV’23],
OneLLM [CVPR’24]



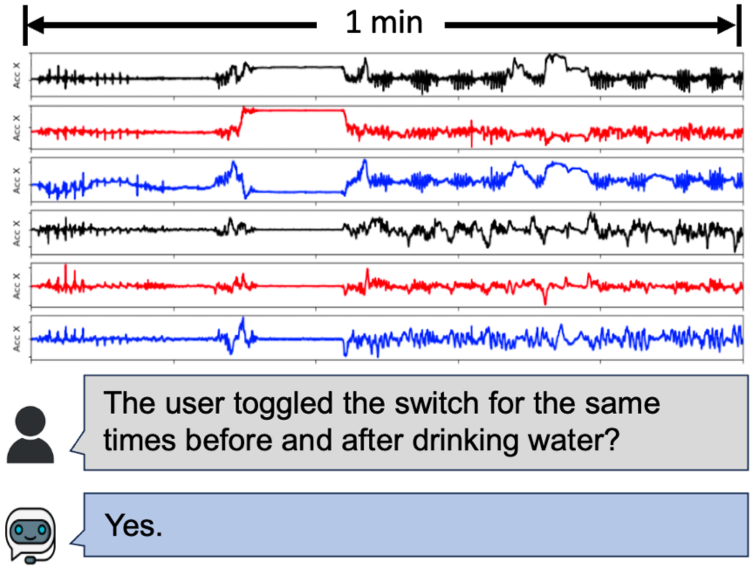

Health-LLM [PMLR’24],
DrHouse [IMWUT’24]

☹ Limited question and answer types

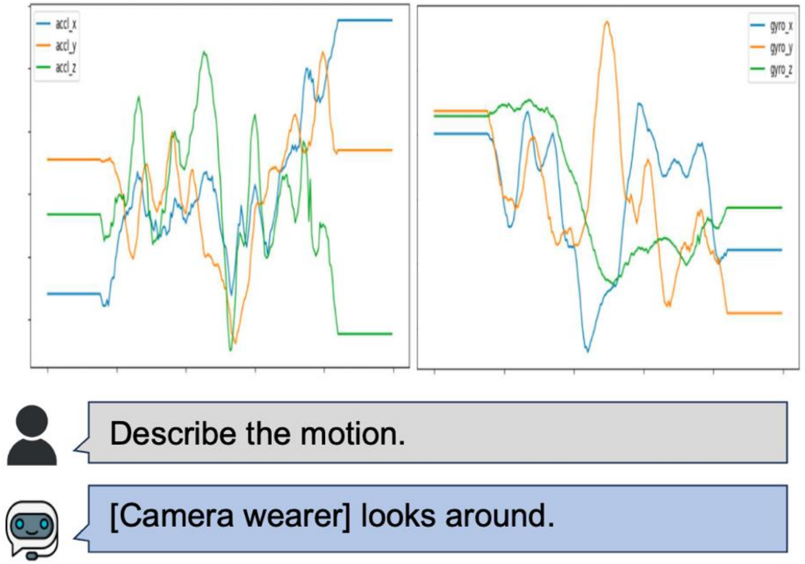
☹ Limited sensor time range

☹ Low-frequency sensor data

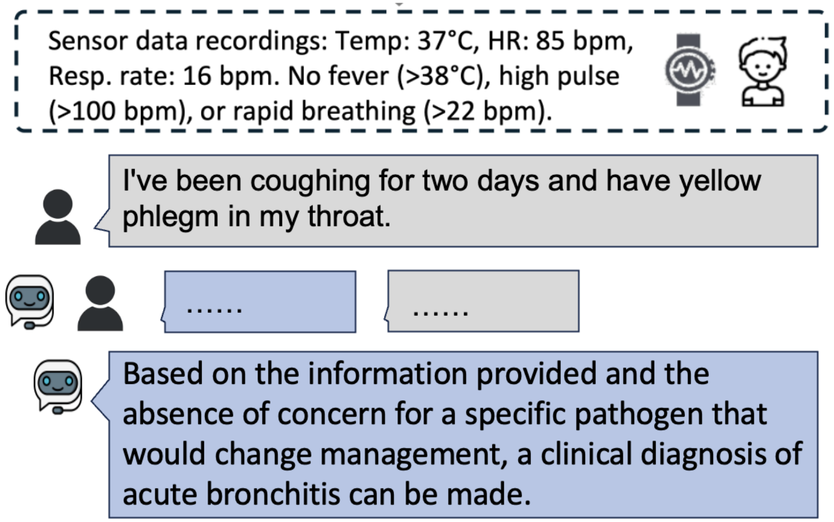
# State of the Art




DeepSQA [IoTDI’21]




AnyMAL [arXiV’23],
OneLLM [CVPR’24]




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

No existing QA benchmark has included **practical, diverse** QA and **long-duration** 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

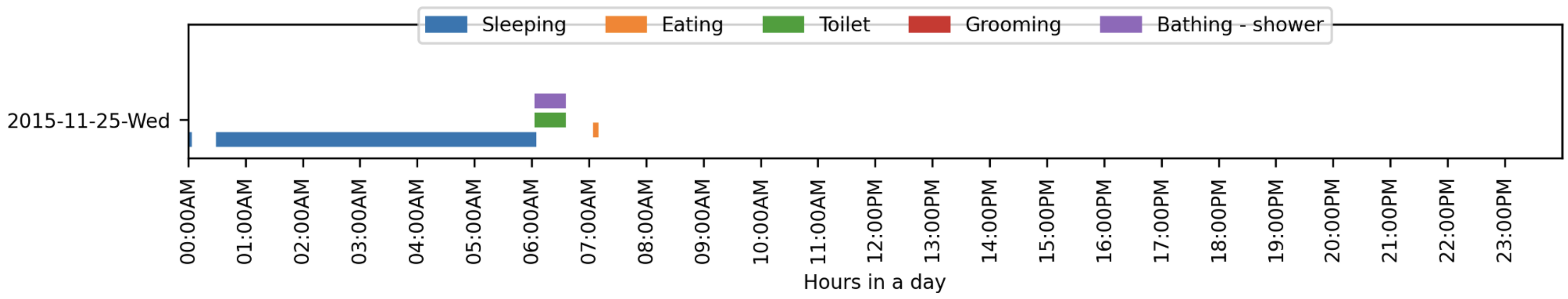
| Goals | SensorQA Design |
| --- | --- |
| • Naturally collected sensor data with **long time span** | • Sensor data from ExtraSensory [IMWUT'17]<br>  • IMUs on phone & watch, audio (MFCC), GPS, compass, phone status, etc<br>  • 60 users, 51 activity labels, 2-10 days |
| • **Diverse** questions and answers that align with human interests | • Crowdsourcing Q&A pairs using Amazon Mechanical Turk[1]<br>  • Multi-time scale activity graph<br>  • 14 label subsets on different life aspects |

Reichman* and Yu* et al. SenSys'25

[1] https://www.mturk.com/

# Our Contribution: SensorQA Dataset (Cont.)

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




Input to workers

What did I do right after waking up on Wednesday?

You showered and toileted right after waking up on Wed.

Output from workers

- 3K questions for a single day and 2.6K questions for longer durations up to weeks
- Correctly answering the questions may require multi-step multimodal reasoning and quantitative analysis

# SensorQA Profile

- SensorQA has 6 question categories and 7 answer categories, focusing on diverse life aspects from activities, locations, to work-life balance

| Question Categories | Example Questions | # of Questions |
|---|---|---|
| Time Compare | Did I spend more time sitting or standing? | 1,432 |
| Day Query | On which day did I spend the most time at home? | 1,277 |
| Time Query | How long was I in class and at school? | 1,119 |
| Counting | How often did I groom? | 725 |
| Existence | Did I have a meeting on Wednesday? | 668 |
| Action Query | What did I do after I left home on Tuesday? | 428 |

(a) Question categories.

| Answer Categories | Example Shortened Answers | # of Answers |
|---|---|---|
| Action | Doing computer work | 1,357 |
| Day/Days | Last Friday | 1,242 |
| Existence | Yes/No | 1,047 |
| Time Length | 40 Minutes | 1,018 |
| Location | At school | 792 |
| Count | Three times | 401 |
| Timestamp | Around 11:00 am | 310 |

(b) Answer categories.

**Table 3: Q&A categories in the SensorQA dataset [3]. The short answers are presented for simplicity.**

# Experimental Setup

- Hardware Platform: NVIDIA Jetson TX2
- 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

# Key Results on SensorQA

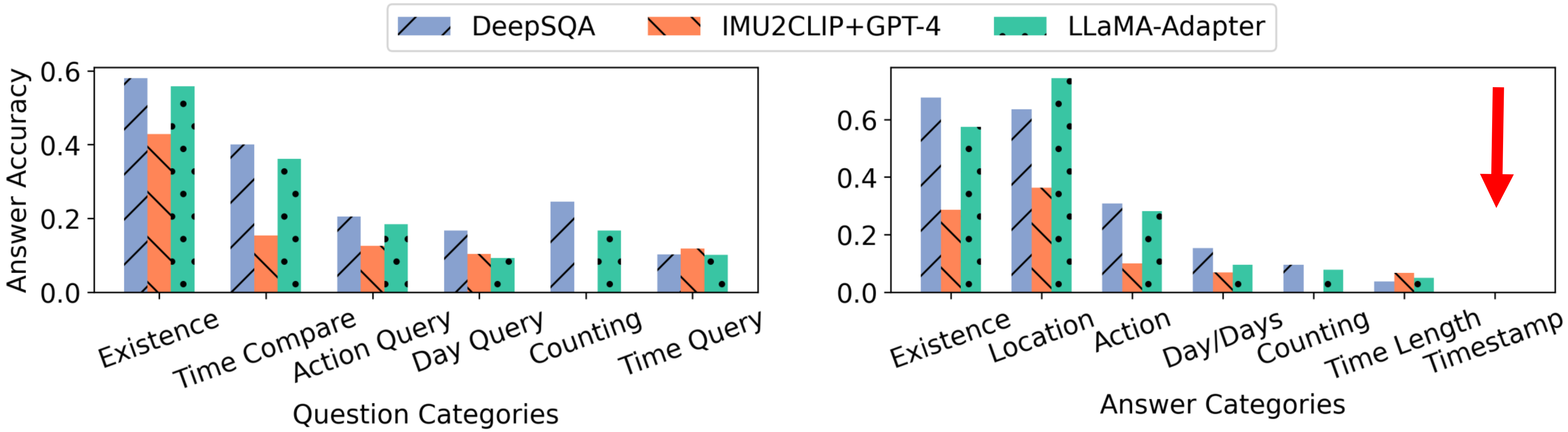
- **Answer accuracy:** matching key phrases in the generated vs. true answers

| 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

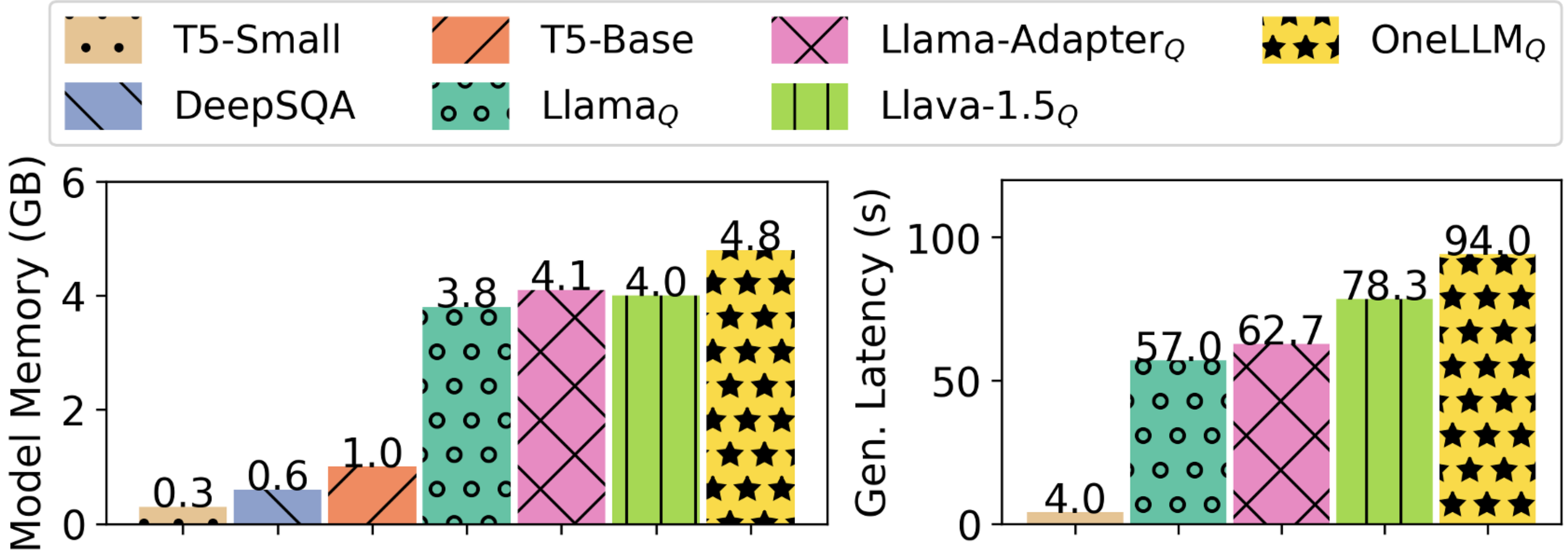
# Key Results on SensorQA (Cont.)

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




**Lesson 2:** SOTA methods struggle with accurate quantitative answers

# Efficiency Results on Jetson TX2




- All LLM-based models are quantized to 4-bit weights using AWQ [MLSys'24]
- LLM-based methods require large memory and have impractical generation latencies of over 57 seconds, highlighting the needs for future improvements

# Conclusion

- Natural language interaction is key to make sensor data more accessible and useful to human users
- Prior benchmarks are limited Q&A types, sensor time range or data complexity
- We introduce SensorQA, the first human-created dataset and benchmark for QA interactions between humans and long0term time-series sensor data
- We benchmark state-of-the-art baselines on SensorQA using typical edge devices, highlighting the challenges in QA accuracy and efficiency
- Dataset and code are available at: https://github.com/benjamin-reichman/SensorQA