SensorQA: A Question Answering Benchmark for Daily-Life Monitoring

Benjamin Reichman*
bzr@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Xiaofan Yu*
x1yu@ucsd.edu
University of California San Diego
La Jolla, California, USA

Lanxiang Hu
lah003@ucsd.edu
University of California San Diego
La Jolla, California, USA

Jack Truxal
jtruxal6@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Atishay Jain
atishay.jain@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Rushil Chandrupatla
ruchandrupatla@ucsd.edu
University of California San Diego
La Jolla, California, USA

Tajana Šimunić Rosing
tajana@ucsd.edu
University of California San Diego
La Jolla, California, USA

Larry Heck
larryheck@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

With the rapid growth in sensor data, effectively interpreting and interfacing with these data in a human-understandable way has become crucial. While existing research primarily focuses on learning classification models, fewer studies have explored how end users can actively extract useful insights from sensor data, often hindered by the lack of a proper dataset. To address this gap, we introduce SensorQA, the first human-created question-answering (QA) dataset for daily life monitoring, based on long-term time-series sensor data. SensorQA is created by human workers and includes 5.6K diverse and practical queries that reflect genuine human interests, paired with accurate answers derived from the sensor data. We further establish benchmarks for state-of-the-art AI models on this dataset and evaluate their performance on typical edge devices. Our results reveal a gap between current models and optimal QA performance as well as efficiency, highlighting the need for new contributions. The dataset and code are available at: https://github.com/benjamin-reichman/SensorQA.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Question Answering, Multimodal Sensors, LLM

*Both authors contributed equally to this research.

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

ACM ISBN 979-8-4007-1479-5/2025/05
https://doi.org/10.1145/3715014.3722074

ACM Reference Format:
Benjamin Reichman*, Xiaofan Yu*, Lanxiang Hu, Jack Truxal, Atishay Jain, Rushil Chandrupatla, Tajana Šimunić Rosing, and Larry Heck. 2025. SensorQA: A Question Answering Benchmark for Daily-Life Monitoring. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3715014.3722074

1 INTRODUCTION

In recent years, the number of connected Internet-of-Things (IoT) devices has grown exponentially, with an estimated 40 billion devices expected by 2030 [43]. These devices generate vast amounts of sensor data, which, unlike text or video, are not easily interpretable by humans due to their raw and complex nature. While existing machine learning algorithms can classify sensor data into predefined categories [7, 34, 35, 49, 51, 52], they fall short in providing an intuitive way for humans to interact with and extract meaningful insights from this data. For example, answering a question like "How good was my work-life balance last week?" is straightforward for humans, but current technologies require multiple steps such as: (1) selecting the appropriate sensor data, (2) understanding the difference between work and relaxing activities, and (3) researching online to understand what qualifies as a healthy work-life balance. In everyday life, people are more interested in gaining insights related to their health and well-being, rather than identifying specific activities at a given moment.

Question Answering (QA) is an ideal framework for modeling natural interactions between humans and sensors: users ask questions and receive accurate answers based on the sensor data. While QA has been extensively studied in the language and vision domains [6, 11, 14, 21, 37, 40, 41], few studies have explored sensor applications using available sensor data. Early QA systems such as *AI therapist* [32] and DeepSQA [50] focus on mental health diagnosis and human activity monitoring. However, their QA dataset is generated by template-based searches, making it limited in diversity and practical value. The recent rise of Large Language Models (LLMs) offers the potential for handling more diverse and sophisticated

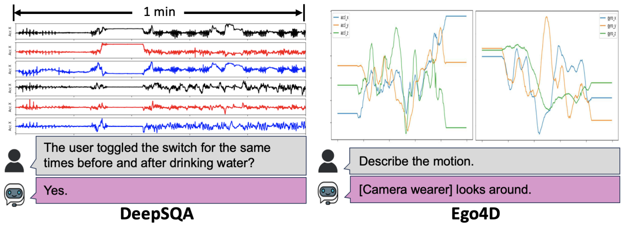


Figure 1: Visualizations of existing QA datasets, DeepSQA [50] and Ego4D [15], using time series data from IMU sensors.

queries, such as captioning IMU signals related to activity [17, 30] or analyzing smartwatch data for medical diagnosis [13, 22, 53]. However, these models are currently constrained to dealing with short durations of sensor data, typically around 10 seconds, or low-dimensional sensor data such as step counts per day.

In summary, the progress of QA interactions for sensing applications is limited by *the absence of suitable datasets and benchmarks*. An ideal dataset would include diverse, practical samples, such as raw sensor signals of varying durations collected in real-world settings, along with rich QA pairs that align with users' genuine interests. To the best of the authors' knowledge, no such benchmark has been proposed.

Our Contribution In this work, we introduce SensorQA, the *first* human-created dataset and benchmark for QA interactions between humans and long-term time-series sensor data. The creation of SensorQA emphasizes realistic QA scenarios that closely resemble everyday life. On the sensor data side, SensorQA builds on the large-scale Extrasensory dataset [46, 47], focusing on daily activity monitoring with commonly available mobile devices, i.e., smartphones and smartwatches. The dataset includes extensive sensor data collected from 60 users intermittently over a period of up to three months. For the QA part, we visualize daily activities in Extrasensory [46, 47] as graphs and present these graphs to human workers on the Amazon Mechanical Turk (AMT) platform. Workers are tasked with generating questions based on their practical interests and writing down ground-truth answers according to the activity graphs. To encourage question diversity, we design multi-timescale activity graphs using 14 different activity label subsets. As a result, SensorQA contains 5.6K QA pairs covering different sensor time scales from one day to multiple weeks, with six question categories and seven answer categories.

The contributions of this work are summarized as follows:

- We present SensorQA, a human-created QA benchmark with naturally collected sensor data and diverse QA pairs, aimed at real-world scenarios.
- We benchmark state-of-the-art baselines on SensorQA using typical edge devices, highlighting the challenges in QA accuracy and efficiency.
- We open-sourced SensorQA and our code to encourage further contributions in this area.

2 RELATED WORK

Question Answering using sensor data Question Answering has attracted extensive interests in the field of Natural Language Processing (NLP), with a wide range of datasets designed to address various language tasks [6, 9, 11, 20, 21, 23, 27, 40, 55, 58]. Recent benchmarks integrate multiple modalities in QA, such as VQAv2 [14] for visual question answering, ScienceQA [29] for science-related questions, and PathVQA [18] for pathology image-based questions, among others. In sensing, researchers have developed multiple QA benchmarks for remote sensing [28, 42, 44, 48] and clinical diagnosis using low-dimensional sensor data [13, 22, 53].

Table 1: Comparing SensorQA and existing QA benchmarks.

Dataset/Benchmark	Human-created rich text	Long-duration sensor data
DeepSQA [50]	×	×
AnyMAL [30], OneLLM [17]	✓	×
SensorQA (this work)	✓	✓

DeepSQA [50] is currently the only QA dataset and benchmark for time-series data from IoT sensors, specifically for human activity monitoring. The dataset is generated automatically by a rule-based search algorithm. As a result, the questions in DeepSQA lack linguistic and content diversity, and, more importantly, may not reflect the practical interests of users. For example, as shown in Fig. 1 (left), one predefined question template compares the frequency of activities at two different times, which may not provide meaningful insights in real-world scenarios.

Multimodal reasoning using sensor data Recent works have explored multimodal reasoning that connects sensor data with natural language. IMU2CLIP [31], mmCLIP [7], and TENT [59] align textual and sensor data embeddings using contrastive learning, similar to CLIP [38]. FM-Fi [49] leverages vision-based models for radio-frequency sensing. The most relevant works, AnyMAL [30] and OneLLM [17], enable more advanced reasoning beyond activity classification. Both works connect sensor embeddings to an LLM via an adapter module, with the LLM fine-tuned on IMU data and text descriptions from the Ego4D dataset [15], as shown in Fig. 1 (right). However, all these methods are limited to simple reasoning tasks over short, fixed-duration signals.

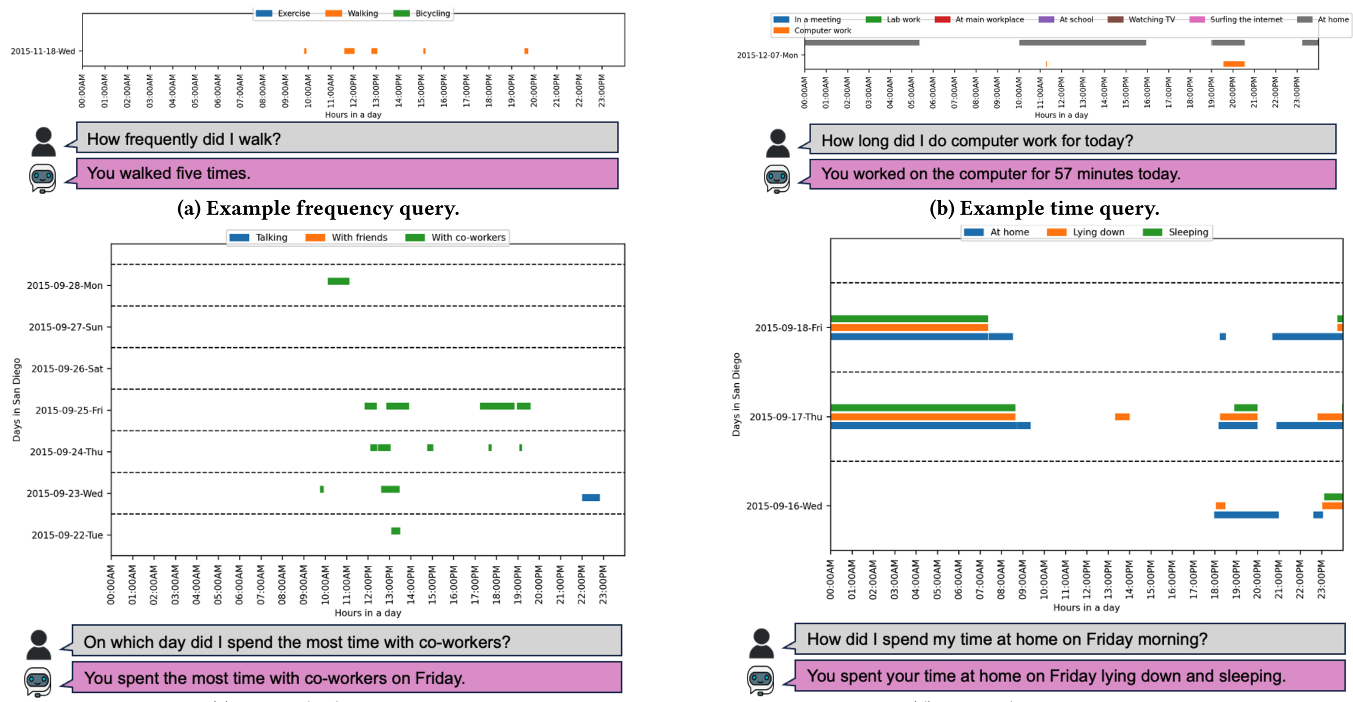
In summary, our SensorQA dataset significantly differs from prior benchmarks in two aspects: (i) On the *QA* side, SensorQA, created by humans, is highly diverse in terms of both the questions posed and the answers that are provided, better reflecting the needs of a user, (ii) On the *sensor* side, SensorQA contains arbitrary length sensor signals and activity histories of up to multiple weeks. The comparison is detailed in Table 1.

3 SENSORQA DATASET

In this section, we describe the collection process for the SensorQA dataset. The real-world scenario that SensorQA models is a long-term, in-the-wild sensor data collection scenario where users follow their daily routines without needing to focus on the sensing device. Throughout this process, users may pose arbitrary questions of personal interest about sensor data, whether focusing on a single day or over several weeks, and expect accurate answers derived from the collected data. To capture such real-world scenarios, we carefully design and implement protocols for both the sensor data (Sec. 3.1) and the QA data (Sec. 3.2) collection for developing SensorQA.

3.1 Sensor Data Collection

We choose to utilize a pre-existing dataset, the ExtraSensory dataset [46, 47] as the source of the sensor data for SensorQA. We select ExtraSensory for its natural, in-the-wild collection setting



(a) Example frequency query. **(b) Example time query.**

(c) Example day query. **(d) Example activity query.**

Figure 2: Example QA pairs in SensorQA. (a) and (b) are generated from daily graphs, while (c) and (d) are generated from multi-day graphs.

and its massive scale. In contrast to the vast majority of sensor datasets [8, 35, 54] that are collected in a heavily controlled environment, ExtraSensory emphasizes real-life settings. This dataset uses easily accessible sensors (IMU, compass, location, audio, and phone state sensors on smartphones and smartwatches), with participants encouraged to maintain their natural routines throughout the collection period ranging from two days to three months. In contrast to Ego4D [15] which relies on data captured by a head-mounted camera, ExtraSensory imposes no restrictions on device placement, whether they be on desks, in pockets, or held in hand. These real-life collection protocols make ExtraSensory an ideal base for the SensorQA dataset. Moreover, ExtraSensory contains raw sensor measurements from 60 subjects and has more than 300K minutes of data, tagged by 51 activity or context labels after cleaning. The massive scale of ExtraSensory allows us to explore a wide range of real-life activities and personal routines during QA. We also note that the data collection protocol for SensorQA can be extended to *any* future sensor application.

3.2 QA Data Collection

We use Amazon Mechanical Turk (AMT) [1], a crowdsourcing platform, to generate question-answer pairs through paid tasks completed by human workers. AMT is extensively utilized in the field of NLP for dataset creation [14, 41]. Unlike template-based question and answer generation in [50] or simply using the narration text provided in the sensor dataset [17, 30], our approach leverages the unique value of human-generated content, ensuring all QA pairs reflect genuine human interests and needs.

We carefully design our QA data collection process to ensure *practical* and *diverse* QA generation. Our objective is to produce a high-quality QA dataset that encompasses a variety of Q&A types and time durations, thereby challenging the AI agent to perform effectively in real-world scenarios. To achieve this, we use two key strategies: (i) we develop multi-time scale activity graphs to facilitate the generation of both short-term and long-term questions, and (ii) we divide the context labels from the ExtraSensory dataset into subsets to encourage queries covering a wide range of aspects. Next, we detail each strategy.

Creating multi-time scale activity graphs Collecting a dataset on AMT requires visualizing the sensor data for the workers and asking them to generate relevant Q&A pairs. However, visualizing raw sensor signals presents a unique challenge due to the inherent unreadability of sensor data by humans [50]. Given our primary interest in understanding the underlying daily activities, we opt to visualize the activity or context labels over time and provide these graphs to the workers. These visual representations, depicted in Fig. 2, resemble Gantt Charts, with the x-axis showing wall-clock time from 00:00 AM to 23:59 PM and the y-axis representing separate days. Different activity labels are shown by bars with distinct colors. These graphs offer an intuitive visualization of daily activities along with the specific timestamps.

While the temporal scale of questions is determined by individual crowdsourcing workers, we recognize that the time scale depicted in activity graphs can implicitly influence their approach. For instance, when presented with a weekly activity graph, workers tend to ask more high-level and qualitative questions like "How frequently do I exercise?" rather than basic quantitative inquiries such as "What did I do at 10:00 AM?" Motivated by this understanding, we have developed graphs at different temporal granularities to prompt questions across various scales:

Label Subset Name	Labels in Subset	Focus of Subset
Main Activity	Lying down, Sitting, Standing, Walking, Bicycling	Posture pattern
Location	Indoor, Outside, At home, At main workplace, At school, In class	Location track
Dietary	Walking, Bicycling, Eating, Exercising	Eating and exercising patterns
Eat	Cooking, Eating, Cleaning	Eating patterns
Sleep	At home, Lying down, Sleeping	Sleeping patterns
Basic Needs	Sleeping, Eating, Grooming, Bathing, In Toilet	Physiological needs
Cleaning	Cleaning, Cooking, Grooming	Cleaning-related activities
Commute	In a vehicle, Walking, Outside, At home	Commute patterns
Exercise	Exercising, Walking, Bicycling	Exercising patterns
Electronics	Watching TV, Surfing the Internet, Phone in Hand, Computer Work	Activities using electronic devices
Social	Talking, With Friends, With Co-Workers	Activities with other people
Work	Sitting, Standing, Computer Work, Lab Work, In a meeting, In class, At school, At main workplace	Work-related activities and locations
Student	Computer Work, Lab Work, In class, At school	Study-related activities and locations
Work-Life Balance	Computer Work, Lab Work, In a meeting, At main workplace, At school, Watching TV, Surfing the internet, At home	Work and relaxation patterns

Table 2: Label subsets used for organizing the graphs and encouraging focus on a specific aspect of daily life.

- **Daily graph with timetable** We generate activity graphs for each user on a single day, accompanied by a table listing the start time, end time, and duration of all activities occurring on that day. This daily graph prompts workers to generate *basic quantitative questions* about specific times and activities on a given day, as shown in Fig. 2a and 2b, where we omit the detailed timetable due to space limitation.
- **Multi-day graph** We also create graphs depicting a user's activities over multiple days. The multi-day graph encourages workers to focus on the general and high-level activity patterns, leading to the generation of *qualitative reasoning questions* as exemplified in Fig. 2c and 2d.

Using multi-time scale activity graphs effectively balances the temporal scale of our collected dataset.

Creating label subsets The labels displayed in each activity graph guide workers to focus on specific aspects of daily life. Therefore, we carefully group these labels into subsets to cover comprehensive and diverse life aspects that users may want to monitor. We create 14 subsets of labels as listed in Table 2, covering everything from essential living needs to work to social activities. Note, that we exclude the labels that contain relatively fewer samples to mitigate the negative impact of the imbalance class distribution in ExtraSensory. During SensorQA collection, we visualize one subset per graph and generate an equal number of questions per label subset. This approach ensures the diversity of questions and answers, enhancing the practicality and difficulty of our dataset.

QA data collection We released the graphs and conducted the QA data collection on AMT over a three-week period. To maintain a balance between quantitative and qualitative questions, we requested for one question per daily graph and three questions per multi-day graph. To make the process more practical, workers were instructed to role-play as if the data were from their own wearable devices. They were asked to create questions from a first-person perspective, based on what would genuinely interest them, and to provide answers from a second-person perspective. The instructions were as follows: *"Pretend you have a smart device with access to the information in the graph. Look at the graph and create a first-person question that requires information from the graph and would interest someone with a smart device. Then, answer in the second person, using the provided examples as a guide."* We then offered six QA examples, generated by the authors, for the workers to reference.

4 DATASET ANALYSIS

In this section, we provide quantitative and qualitative analysis of the SensorQA to better understand its characteristics.

Examples of the collected Q&As in SensorQA are displayed in Fig. 2. SensorQA contains 5,648 question-answer pairs, with an average length of 10.43 words per question and 10.48 words per answer. The dataset has a total of 118,051 tokens, of which 1,709 are unique and primarily related to daily activities. The repetition of words makes it more challenging for AI agents to answer questions accurately, as they must differentiate between similar questions based on the specifics of the sensor data.

To closely inspect the diversity of SensorQA, we profile the question and answer categories. We manually label 200 pairs, then train two BERT models [12] to classify the question and answer categories, respectively. The final profiling results are displayed in Table 3. SensorQA includes six distinct question categories and seven answer categories. The distribution of questions and answers is imbalanced, with a notable focus on time-related aspects of activities, as seen in the high number of questions in the "Time Compare" and "Time Query" categories. This pattern aligns with practical user interests but has not been observed in previous QA datasets for human activities [30, 31, 50]. In addition to time-related queries, SensorQA covers a wide range of other aspects, including action, location, counting, and existence, demonstrating its diversity and practicality.

5 BENCHMARK RESULTS

In this section, we benchmark state-of-the-art AI models on the SensorQA dataset and reveal the gap between existing models and ideal performance.

5.1 Benchmark Setup

We establish comprehensive baselines using three distinct modality combinations: text-only, vision+text, and sensor+text, to identify the impact of each modality. We use few-shot learning (FSL) for closed-source models like GPT, and apply LoRA fine-tuning (FT) [19] for open-source models like Llama. We randomly split 80% of the data in SensorQA for training and 20% for testing. More details are included in the github repository[1].

[1]https://github.com/benjamin-reichman/SensorQA?tab=readme-ov-file

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.

Baselines We begin by evaluating the generative pretrained models using few-shot QA examples in the prompt.

- **GPT-3.5-Turbo [56] and GPT-4 [4]** are text-only baselines that only use the questions as input.
- **GPT-4-Turbo [4] and GPT-4o [4]** are vision+text baselines that use both the questions and activity graphs.
- **IMU2CLIP-GPT4 [31]** is the state-of-the-art sensor+text GPT baseline. It first uses a trained CLIP model to retrieve the most relevant text for each chunk of IMU signal, then combines the text into a storyline and provides it to GPT-4, along with the question, for answer generation. We train the CLIP model using ExtraSensory [46, 47].

We also selected the following open-source models, which are tested after they are either trained or finetuned. We started from the official code release and used the default hyperparameters. All Llama-based backbones are Llama-2 7B [45] unless specified otherwise.

- **T5 [39]** and **Llama [45]** are widely used language models, serving as text-only baselines.
- **Llama-Adapter [57]** is a vision+text framework that combines vision inputs (activity graphs) with a Llama model via a pretrained transformer-based adapter.
- **Llava-1.5 [26]** is a state-of-the-art vision+text model that integrates a visual encoder and Vicuna [10] for visual and language understanding.
- **DeepSQA-CA [50]** fuses sensor+text modalities by training a CNN-LSTM model with compositional attention, for predicting from a limited set of answers given questions and IMU signals.
- **OneLLM [16]** is a state-of-the-art multimodal LLM framework that supports eight modalities, including sensor+text data. It feeds sensor data through a pretrained CLIP encoder and uses a mixture of projection experts for modality alignment.

Metrics We consider two versions of SensorQA. For the full-answer version, we use commonly applied NLP metrics, Rouge [24], Meteor [5] and Bleu [36] scores. These scores assess n-gram precision between the generated and ground-truth answers. Intuitively, higher scores indicate more overlaps. We further distill a short answer version of SensorQA by prompting GPT-3.5-Turbo [56] to extract the 1-2 keywords from each full answer. We then use exact-match accuracy, i.e., whether the keywords appear in the generated

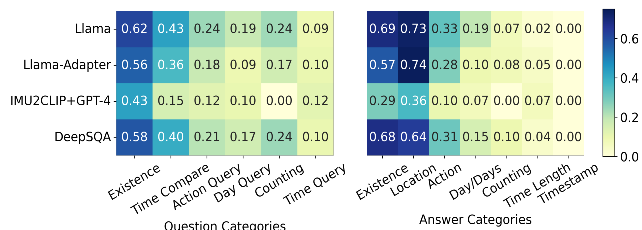


Figure 3: Exact-match accuracy of representative baselines displayed by question (left) and answer category (right).

answers, to evaluate the short answers. The two versions of SensorQA provide a comprehensive evaluation of both the qualitative and quantitative aspects of the generated answers.

For efficiency, we aim for baseline models to ultimately be deployable on edge devices as personal AI assistants. Therefore, we evaluate their memory requirements and average generating latency per answer on the NVIDIA Jetson TX2 [2], a typical edge platform featuring an NVIDIA Pascal GPU with 256 CUDA cores and 8GB of RAM. Here, all Llama-based models are quantized to 4-bit weights using AWQ [25].

5.2 Results

QA performance Table 4 presents the results of all baselines on both the full-answer and short-answer versions of SensorQA. The Rouge, Meteor, and Bleu scores focus on the language quality of full answers, while exact-match accuracy evaluates the factual correctness of answers. The results show that existing AI models perform poorly on SensorQA, with the best-performing baseline Llama-Adapter achieving an accuracy of only 28%. This highlights the significant challenge SensorQA poses for current models. Specifically, our exact-match accuracy metric is a stringent yet realistic measure of QA correctness. For instance, "10 minutes" versus "20 minutes" is considered incorrect. Answering questions in SensorQA is especially challenging due to the broad range of Q&A categories and an extensive word corpus to choose from.

In Table 4, the Sensor+Text baselines perform worse than even the Text-only baselines. This is because most existing models, such as DeepSQA [50] and OneLLM [17], are optimized for fusing short-duration sensor data with natural language. However, when applied to SensorQA's long-duration sensor data, these models fail to effectively fuse sensor and text data, resulting in poorer performance than using text alone. These results emphasize the need for new approaches to effectively integrate long-term sensor data and text in realistic applications like SensorQA.

Modalities	Backbone Model	FSL/FT[1]	Full Answers					Short Answers
			Rouge-1 (↑)	Rouge-2 (↑)	Rouge-L (↑)	Meteor (↑)	Bleu (↑)	Accuracy (↑)
Text	GPT-3.5-Turbo	FSL	0.35	0.23	0.32	0.43	0.16	3.0%
Text	GPT-4	FSL	0.66	0.51	0.64	0.66	<u>0.39</u>	16.0%
Text	T5-Base	FT	0.71	0.55	0.69	<u>0.70</u>	**0.43**	25.4%
Text	Llama	FT	<u>0.72</u>	**0.62**	**0.72**	**0.72**	0.38	26.5%
Vision+Text	GPT-4-Turbo	FSL	0.38	0.28	0.36	0.51	0.15	14.0%
Vision+Text	GPT-4o	FSL	0.39	0.28	0.37	0.61	0.25	7.0%
Vision+Text	Llama-Adapter	FT	**0.73**	<u>0.57</u>	<u>0.71</u>	**0.72**	**0.43**	**28.0%**
Vision+Text	Llava-1.5	FT	0.62	0.46	0.60	0.58	0.35	21.5%
Sensor+Text	IMU2CLIP-GPT4	FSL	0.44	0.28	0.40	0.53	0.16	13.0%
Sensor+Text	DeepSQA	FT	0.34	0.05	0.34	0.18	0.0	<u>27.4%</u>
Sensor+Text	OneLLM	FT	0.12	0.04	0.12	0.04	0.0	5.0%

[1]FS: Few-Shot Learning. FT: Finetuning.

Table 4: Benchmark results of baselines on SensorQA. Bold and underlined values show the best and second-best results.

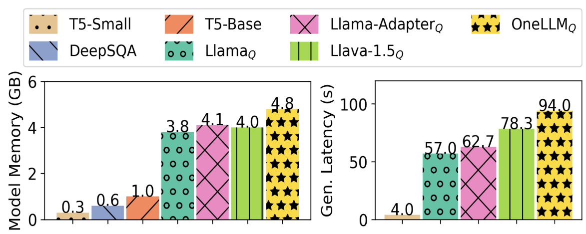


Figure 4: Model memory size (left) and average answer generating latency (right) on Jetson TX2 [2]. Footnote Q denotes models after quantization.

Overall, finetuning open-source models outperforms few-shot learning on GPT baselines. For instance, Llama-Adapter achieves 0% exact-match accuracy without fine-tuning, highlighting the importance of fine-tuning on the target dataset to better adapt models for specific tasks.

QA performance per category Different Q&A types present varying difficulties for the models. Fig 3 shows exact-match accuracy by category using four representative baselines. Existence questions achieve the highest accuracy, since they require only a Yes/No response. However, even the best baseline performs only marginally better than random guessing, with an accuracy of 62%. The most challenging category involves time-related queries, such as duration and timestamp questions, which is one unique property of SensorQA compared to prior datasets [30, 31, 50]. It is critical for future approaches to accurately extract time information from sensor data and incorporate it into responses.

Efficiency results Fig. 4 presents model memory requirements and average answer generation latency on the NVIDIA Jetson TX2 [2]. DeepSQA and T5-Base encounter out-of-memory (OOM) issues with large multi-day timeseries inputs, so their latency results are omitted. Non-LLM models require less memory and shorter latency than LLM-based models. However, non-LLM methods also show poor QA performance. LLM-based models, though more accurate, require large memory for their billions of parameters and have impractical generation latencies of over 57 seconds, even after quantization. Multimodal LLMs experience additional delays due to the need to encode image or sensor data before LLM inference. Optimizing memory and efficiency are crucial challenges in developing future conversational AI for mobile deployment.

6 DISCUSSION

Imbalanced quries As shown in Table 3, SensorQA exhibits a skewed distribution towards various question and answer categories, with a particular emphasis on time-related queries. We recognize the skewed time-related aspects in SensorQA as a valuable characteristic. In practice, time-related information, particularly the durations of specific activities, provides critical insights into a user's lifestyle and health [33]. SensorQA captures this trend by asking turkers to imagine themselves as the owners of sensing devices and compose questions that align with their interests. Therefore, we see the emphasis on time-related queries in SensorQA as a feature that reflects genuine user needs and interests.

Limitations of SensorQA The SensorQA is based on ExtraSensory [46, 47] and shares the biases and limitations of this dataset. The activity label at times was restrictive and may have constrained the variety of possible questions asked. Future work could expand the size of SensorQA or explore additional label subsets for activity graph generation. We also recognize the opportunity in incorporating subjective metrics defined by the users, which offers a more comprehensive evaluation beyond a rigid accuracy metric. Last but not least, the methodology of creating SensorQA can be seamlessly extended to other sensor applications in the future.

7 CONCLUSION

As IoT and wearable devices proliferate, the ability to interact with sensor data through conversational AI becomes increasingly crucial. In this work, we introduce SensorQA, a question-answering dataset created by humans to foster natural language interactions between humans and wearable sensors in daily life monitoring. SensorQA is built on sensor data from ExtraSensory [46, 47] and 5.6K QA pairs collected from AMT, featuring practical scenarios and diverse queries. Benchmarking results on state-of-the-art AI models demonstrate the gap between existing solutions and ideal performance, both in QA and efficiency.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation under Grants #2003279, #1826967, #2100237, #2112167, #1911095, #2112665, #2120019, #2211386 and in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA.

REFERENCES

[1] 2024. Amazon Mechanical Turk. https://www.mturk.com/. [Online].

[2] 2024. Jetson TX2 Module. https://developer.nvidia.com/embedded/jetson-tx2. [Online].

[3] 2024. SensorQA Dataset. https://anonymous.4open.science/r/SensorQA-373E/. [Online].

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

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* 65–72.

[6] Asma Ben Abacha and Dina Demner-Fushman. 2019. A Question-Entailment Approach to Question Answering. *BMC Bioinform.* 20, 1 (2019), 511:1–511:23. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3119-4

[7] Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 2024. mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems.* 184–197.

[8] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042.

[9] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3697–3711. https://doi.org/10.18653/v1/2021.emnlp-main.300

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[11] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2924–2936. https://doi.org/10.18653/v1/N19-1300

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Chun-Cheng Chang, Xuhai" Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18995–19012.

[16] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700* (2023).

[17] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[18] Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering. *ArXiv* abs/2003.10286 (2020). https://api.semanticscholar.org/CorpusID:214612106

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[20] Mengkang Hu, Haoyu Dong, Ping Luo, Shi Han, and Dongmei Zhang. 2024. KET-QA: A Dataset for Knowledge Enhanced Table Question Answering. *arXiv preprint arXiv:2405.08099* (2024).

[21] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1601–1611. https://doi.org/10.18653/v1/P17-1147

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

[23] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out.* 74–81.

[25] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. In *MLSys*.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 26296–26306.

[27] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems* 36 (2024).

[28] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 58, 12 (2020), 8555–8566.

[29] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS).*

[30] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058* (2023).

[31] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023.* 13246–13253.

[32] Jingping Nie, Hanya Shao, Minghui Zhao, Stephen Xia, Matthias Preindl, and Xiaofan Jiang. 2022. Conversational ai therapist for daily function screening in home environments. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications.* 31–36.

[33] US Department of Health, Human Services, et al. 2021. Increase the proportion of adults who do enough aerobic physical activity for substantial health benefits—PA-02.

[34] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking.* 324–337.

[35] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services.* 530–543.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 311–318.

[37] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4542–4550.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[40] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna

Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[41] Benjamin Z. Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, Saif Punjwani, Atishay Jain, and Larry Heck. 2023. Outside Knowledge Visual Question Answering Version 2.0. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096074

[42] Argho Sarkar, Tashnim Chowdhury, Robin Murphy, Aryya Gangopadhyay, and Maryam Rahnemoonfar. 2023. Sam-vqa: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* (2023).

[43] Satyajit Sinha. 2023. State of IoT 2024: Number of connected IoT devices growing 13% to 18.8 billion globally. https://iot-analytics.com/number-connected-iot-devices/. [Online].

[44] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. 2023. SatBird: a Dataset for Bird Species Distribution Modeling using Remote Sensing and Citizen Science Data. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=Vn5qZGxGj3

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

[46] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing* 16, 4 (2017), 62–74.

[47] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–22.

[48] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. 2024. EarthVQA: Towards Queryable Earth via Relational Reasoning-Based Remote Sensing Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5481–5489.

[49] Yuxuan Weng, Guoquan Wu, Tianyue Zheng, Yanbing Yang, and Jun Luo. 2024. Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 436–449.

[50] Tianwei Xing, Luis Garcia, Federico Cerutti, Lance Kaplan, Alun Preece, and Mani Srivastava. 2021. DeepSQA: Understanding Sensor Data via Question Answering. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 106–118.

[51] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[52] Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Chen Jiming. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*.

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

[54] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2024. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems* 36 (2024).

[55] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[56] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420* (2023).

[57] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).

[58] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2019. JEC-QA: A Legal-Domain Question Answering Dataset. arXiv:1911.12011 [cs.CL]

[59] Yunjiao Zhou, Jianfei Yang, Han Zou, and Lihua Xie. 2023. TENT: Connect Language Models with IoT Sensors for Zero-Shot Activity Recognition. *arXiv preprint arXiv:2311.08245* (2023).