HARGPT: Are LLMs Zero-Shot Human Activity Recognizers?

Sijie Ji*, Xinzhe Zheng*, Chenshu Wu
Department of Computer Science, The University of Hong Kong
{sijieji, xzzheng, chenshu}@cs.hku.hk

***Abstract*—There is an ongoing debate regarding the potential of Large Language Models (LLMs) as foundational models seamlessly integrated with Cyber-Physical Systems (CPS) for interpreting the physical world. In this paper, we carry out a case study to answer the following question: Are LLMs capable of zero-shot human activity recognition (HAR)? Our study, HARGPT, presents an affirmative answer by demonstrating that LLMs can comprehend raw IMU data and perform HAR tasks in a zero-shot manner, with only appropriate prompts. HARGPT inputs raw IMU data into LLMs and utilizes the *role-play* and *"think step-by-step"* strategies for prompting. We benchmark HARGPT on GPT4 using two public datasets of different inter-class similarities and compare various baselines both based on traditional machine learning and state-of-the-art deep classification models. Remarkably, LLMs successfully recognize human activities from raw IMU data and consistently outperform all the baselines on both datasets. Our findings indicate that by effective prompting, LLMs can interpret raw IMU data based on their knowledge base, possessing a promising potential to analyze raw sensor data of the physical world effectively.**

***Index Terms*—Large Language Models, Cyber-Physical Systems, IoT, Artificial Intelligence**

I. Introduction

Recent advancements in Foundation Models (FMs), particularly Large Language Models (LLMs) and Large Multimodal Models (LMMs), have garnered significant attention due to their remarkable capabilities in efficiently handling a diverse array of downstream tasks. Foundation models are noteworthy for two primary reasons: their ability to mimic human reasoning at a high level and their exceptional generalization abilities. Most importantly, some evidence shows that LLMs have an emergent ability to understand the physical world in a manner akin to that of a child. This is exemplified by Sora [1], which exhibits intuitive comprehension of various aspects of the world, ranging from fundamental physics principles to complex societal and artistic concepts.

Despite these impressive feats, some researchers posit that current LLMs possess only surface-level knowledge and are still far from achieving a deep understanding of the physical world [2]. This limitation is attributed to their training on vast amounts of internet-scale text corpora and images, leading to subpar performance when analyzing digital and time series data. Additionally, these models often struggle when attempting to interact with the physical world in a meaningful way, such as generating precise control sequences [3].

*These authors contributed equally to this work

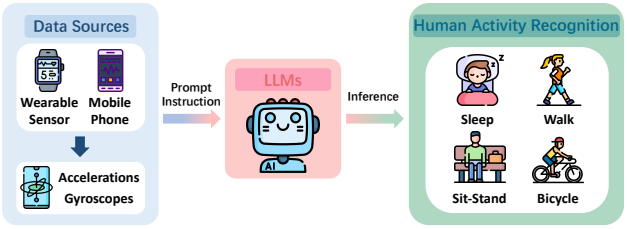


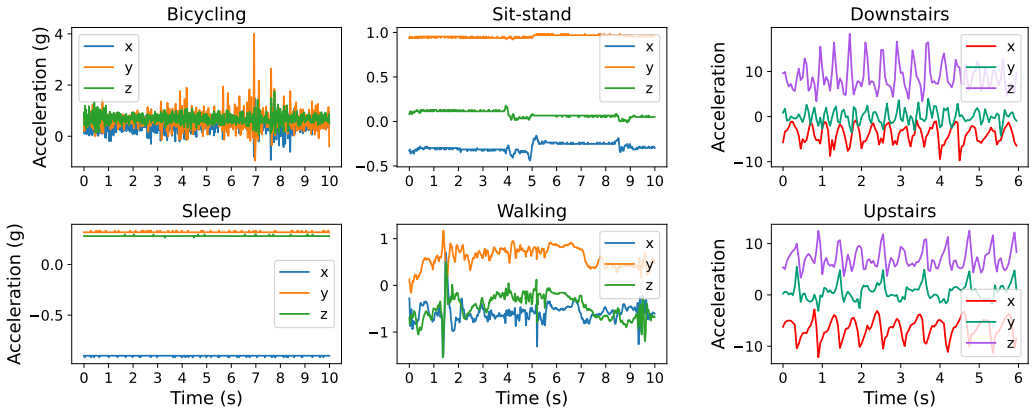
Figure 1: Workflow of HARGPT.

This ongoing debate has sparked considerable interest in exploring the potential of LLMs to serve as foundational models seamlessly integrated with Cyber-Physical Systems (CPS) for interpreting the physical world [4]. In particular, researchers propose the concept of Penetrative AI and explore the integration of LLM with the physical world from two levels: textual signals and raw digital signals [5]. However, the cases they show, such as inferring user location through WiFi signals, mainly rely on explicitly direct textural information and they require specific prompt engineering to provide expert knowledge about the usage of sensor data. The true potential of LLMs as a physical world model remains largely unexplored.

Motivated by the advancements of LLMs and the potential of Penetrative AI, this paper starts from the classic CPS application Human Activity Recognition (HAR), aiming to explore and understand the current capabilities of LLMs by directly inputting raw sensor data with a simple prompt. As illustrated in Fig. 1, the raw IoT sensor data is fed directly into various well-known LLMs such as ChatGPT [6], Google Gemini [7], and LLaMA2-70b [8], yielding recognition outcomes. The performance evaluation on two distinct benchmark datasets with two levels of difficulty, one with distinct patterns, such as sleeping and walking, and the other with inter-class similarities, such as climbing upstairs and downstairs. The results of the experiments demonstrate that LLMs are capable of performing zero-shot HAR using raw sensor data, achieving an average accuracy of 80% on unseen data (same activity performed by unseen users), which outperforms existing methods using classical machine learning or deep learning. More importantly, unlike conventional approaches that are prone to performance degradation when confronted with unseen data and often necessitate retraining or fine-tuning for specific datasets, LLMs exhibit a high degree of robustness.

In summary, we present HARGPT and make the following

979-8-3503-6345-6/24/$31.00 ©2024 IEEE
DOI 10.1109/FMSys62467.2024.00011



(a) Capture24 (b) HHAR

Figure 2: IMU data visualization of two datasets. (a): Capture24 dataset contains four HAR categories with distinct patterns; (b): HHAR dataset contains two similar HAR categories.

remarks:

- HARGPT first reveals that LLMs can function as zero-shot human activity recognizers without the need for fine-tuning or domain-specific, expertise-guided prompt engineering.
- HARGPT showcases the proficiency of LLMs in processing IoT sensor data for carrying out tasks in the physical world.

Furthermore, we delve into exploration and discourse on the insights gained and notable discoveries made while utilizing LLMs for handling IoT sensor data, inspiring future research on adopting LLMs for CPS.

II. HARGPT: Zero-shot HAR with LLMs

A. Experiments

To evaluate the capability of LLMs for HAR based on IMU raw data, we conducted two levels of experiments varying in difficulty. The initial set of experiments aimed to determine if LLMs could differentiate between movement types with distinct patterns, such as sleeping and walking, which exhibit obvious dissimilarities. In the subsequent set of experiments, we sought to ascertain whether these models could further discriminate between movements that possess highly similar patterns, such as upstairs and downstairs, which are challenging to differentiate even for human observers.

1) Dataset Setup: To facilitate these experiments, we utilized two distinct datasets: Capture24 [9] and HHAR [10], for the experiments respectively. Below, further details regarding each dataset are provided.

Capture24: Capture24 is prepared for the first experiment. The dataset contains human motion activities in daily living with accelerations only. The data is collected from wrist-watches with a sampling rate of 100Hz. For the experimental task, we only use the IMU data with labels of sleep, walking, bicycle, and sit-stand. The visualization outcome is given in Fig. 2a. These categories of data have significantly distinct characteristics and are relatively easy to distinguish.

To compare LLMs with the baselines that need training data, we partition the dataset by the user into training, validation, test seen, and test unseen datasets at a ratio of 4:1:1:2. Additionally, given the requirement to input raw IMU data as tokens into LLMs, we down-sample the IMU data to 10Hz for LLMs.

HHAR: HHAR is utilized for the second experiment, so we evaluate downstairs and upstairs data exclusively. It contains readings from accelerometers and gyroscopes from 9 users. The data is collected from mobile phones carried by the users around their waists with sampling rates of 100 and 200Hz. We visualize the IMU data in Fig. 2b, which further confirms our assumptions.

HHAR is partitioned by the user into training, validation, test seen, and test unseen datasets at a ratio of 4:1:1:2 as well. The IMU data is down-sampled to 10Hz as tokens for LLMs.

2) Baseline: Classical machine learning models and state-of-the-art deep learning models for HAR are taken as baseline models. These models need to be trained to acquire classification ability.

Random Forest [11]: Random forests (RF) is an ensemble learning method for classification. It operates by constructing a multitude of decision trees at training time. With its specific learning strategy, RF could handle minor classes well in the classification tasks. For implementation, we use the model provided in scikit-learn [12] with its default settings.

SVM [13]: support vector machine (SVM) is a supervised max-margin model with associated learning algorithms that analyzes data for classification and regression. We use the SVM model in scikit-learn with Gaussian kernels.

DCNN [14]: DCNN is composed of convolutional neural networks that can automatically learn features from multichannel time series signals acquired from body-worn sensors for HAR. It can learn complex HAR features, and achieve promising

Method	Test Subject	Evaluation Metric (macro avg.)		
		Precision	Recall	F1-Score
RF	Seen	0.560	0.635	0.580
	Unseen	0.525	0.598	0.555
SVM	Seen	0.463	0.505	0.478
	Unseen	0.535	0.598	0.545
DCNN	Seen	0.615	0.628	0.615
	Unseen	0.595	0.600	0.588
LIMU-LSTM	Seen	0.615	0.628	0.618
	Unseen	0.595	0.588	0.585
GPT4 – DO*	Unseen	0.498	0.468	0.465
GPT4 – CoT*	Unseen	**0.818**	**0.793**	**0.795**

Table I: Overall test results on Capture24 dataset. (DO*: direct output; CoT*: chain-of-thought.)

classification results compared with some traditional methods, like deep belief network.

LIMU-LSTM [15]: LIMU-LSTM is a classification model constructed with long short-term memory (LSTM). Different from other baseline models, LIMU-LSTM analyzes the input features in chronological order.

3) Prompt Structure: LLMs are generally known as excellent few-shot learners, where one can use a text or template known as a prompt to strongly guide the generation to output answers for desired tasks [16]. For example, in a recent study [17], researchers utilize a guiding instruction to release the professional medical domain knowledge of LLMs, achieving state-of-the-art performance across all benchmark datasets. The problem with such prompt engineering is that they need to manually construct the question-and-answer template, which makes the LLMs restricted to specific downstream applications [5]. Further, the study also shows that such a hand-crafted prompting template will go against the Chain-of-Thought property of LLMs and result in bad performance [18]. To circumvent the above problems, we structure our prompt in the simplest way, using role-play instructions and "let's think step-by-step guidance" to directly trigger and demonstrate the original capabilities of LLMs without giving an answer template.

Our prompt design, depicted in Fig. 3, comprises only an instruction and a question. The instruction aims to leverage the expert knowledge of LLMs regarding IMU. The question provides specific details about data collection, down-sampled raw data sequence, and potential categories of human actions. By concluding the question with the phrase "Please make an analysis step by step," we aim to elicit a detailed chain-of-thought (CoT) process from LLMs, as this approach has been proven to enhance the accuracy of their answers in existing literature [16], [19]. As shown in Fig. 5, when we do not restrict the answer template of LLM, it will generate richer text information utilizing the inherent reasoning capability to retrieve the corresponding embedding knowledge.

Label	Metrics	RF	SVM	DCNN	LIMU-LSTM	GPT4 - DO	GPT4 - CoT
Bicycling	Precision	0.80	0.89	0.95	**0.97**	0.53	0.64
	Recall	0.81	0.79	0.77	0.74	0.36	**0.84**
	F1-score	0.81	0.84	**0.85**	0.84	0.43	0.72
Sit-stand	Precision	0.24	0.29	0.32	0.26	0.35	**0.89**
	Recall	0.41	0.46	0.43	0.33	0.32	**0.73**
	F1-score	0.30	0.36	0.37	0.29	0.33	**0.80**
Sleep	Precision	0.61	0.47	0.75	0.80	0.71	**1.00**
	Recall	0.75	0.83	**0.97**	0.96	0.50	0.83
	F1-score	0.67	0.60	0.85	0.88	0.59	**0.91**
Walking	Precision	0.45	0.49	0.36	0.35	0.40	**0.74**
	Recall	0.42	0.31	0.23	0.32	0.69	**0.77**
	F1-score	0.44	0.38	0.28	0.33	0.51	**0.75**

Table II: Performance of each action category on Capture24 dataset.

Prompt Template

Instruction: You are an expert of IMU-based human activity analysis.
Question: The IMU data is collected from **{device name}** attached to the user's **{location}** with a sampling rate of **{freq}**. The IMU data is given in the IMU coordinate frame. The three-axis accelerations and gyroscopes are given below.
Accelerations:
x-axis: {...}, y-axis: {...}, z-axis: {...}
Gyroscopes:
x-axis: {...}, y-axis: {...}, z-axis: {...}
The person's action belongs to one of the following categories: <***category list***>.
Could you please tell me what action the person was doing based on the given information and IMU readings? Please make an analysis step by step.
Response: **{answer}**

Figure 3: Chain-of-thought prompt design for HARGPT.

It should be noted that Capture24 only contains acceleration data, so we exclude gyroscope information during testing. Additionally, we conduct a comparative test to assess the impact of CoT on improving the accuracy of question answering. In the experiment, the last sentence is replaced with "*Please give your answer directly without analysis.*"

B. Evaluation

We select GPT4 [20], the most advanced and powerful LLM currently accessible, to conduct a comprehensive analysis with four other baseline models across two datasets. Additionally, we compare the prompt mode of direct output (DO) without analysis to verify the effectiveness of applying CoT prompt for LLMs to improve prediction accuracy.

Inter-class Difference: We first test GPT4 and other baseline models on the Capture24 dataset, which contains four generally distinct HAR categories, including sleep, walking, sit-stand, and bicycling. The overall testing result is shown in Tab. I. Undoubtedly, leveraging its robust comprehension capabilities and CoT prompt, GPT4 has exhibited exceptional performance across all baseline measures on the unseen set, boasting an average f1-score of 0.795. Specifically, when compared with GPT4-DO, GPT4-CoT demonstrates a noteworthy improvement of 0.33. In contrast, even the best baseline DCNN achieves only a modest level of approximately 0.6. LIMU-LSTM marginally underperforms compared to DCNN. Conversely, the two remaining machine learning methods ex-

Method	Test Subject	Evaluation Metrics (macro avg.)		
		Precision	Recall	F1-Score
RF	Seen	0.935	0.945	0.935
	Unseen	0.330	0.400	0.325
SVM	Seen	0.835	0.680	0.670
	Unseen	0.735	0.665	0.570
DCNN	Seen	0.980	0.985	0.980
	Unseen	0.535	0.505	0.385
LIMU-LSTM	Seen	0.960	0.985	0.975
	Unseen	0.720	0.700	0.700
GPT4 - DO	Unseen	0.555	0.570	0.565
GPT4 - CoT	Unseen	**0.790**	**0.795**	**0.790**

Table III: Overall test results on HHAR dataset.

hibit poor performance. For instance, the classification results of each category are presented in Tab. II, and it becomes evident that GPT4-CoT could achieve a well-balanced performance. Notably, while DCNN and LIMU-LSTM exhibit high accuracy in predicting bicycling and sleep activities, their accuracy significantly diminishes when predicting sit-stand and walking activities. This disparity in accuracy might be attributed to the increased freedom of motion associated with the latter two activities.

Inter-class Similarity: To avoid the issue of the obvious pattern difference in the first experiment, we conducted a comparison challenge test, in which the model is asked to discern between two similar actions, specifically ascending and descending stairs. Fig. 2b visually depicts the results obtained from this investigation. The experimental findings are presented in Tab. III and Tab. IV. Similarly, GPT4-CoT demonstrates outstanding outcomes, exhibiting an average accuracy close to 80%, along with a commendable recall rate and f1-score. On the other hand, the baseline models, while capable of achieving high accuracy on test seen set (where three of the baseline models surpass 90% accuracy), struggle to perform adequately on test unseen samples. Even the top-performing LIMU-LSTM exhibits a nearly 10% decrease when compared to GPT4-CoT on the unseen dataset.

Detailed Inference Example[1]**:** To showcase the proficiency of GPT4 in producing expert-level knowledge and precise inference, we present a comprehensive illustration of the concept of walking, as depicted in Fig. 5. The inference process is divided into four parts. First of all, it explains the information given and makes it clear what problem it needs to solve. The second step combines hidden "expert knowledge" to accurately characterize the raw IMU data corresponding to each action. For example, bicycling and walking both have periodic characteristics, sleep is almost non-fluctuating, and sit-stand has sudden changes in value. In the third installment,

[1]Due to the space limit, please visit our project page to explore more examples: https://github.com/aiot-lab/HARGPT.

Label	Metrics	RF	SVM	DCNN	LIMU-LSTM	GPT4 - DO	GPT4 - CoT
Downstairs	Precision	0.42	**1.00**	0.55	0.76	0.48	0.73
	Recall	0.72	0.33	0.05	0.57	0.55	**0.80**
	F1-score	0.53	0.50	0.09	0.65	0.51	**0.76**
Upstairs	Precision	0.24	0.47	0.52	0.68	0.63	**0.85**
	Recall	0.08	**1.00**	0.96	0.83	0.59	0.79
	F1-score	0.12	0.64	0.68	0.75	0.62	**0.82**

Table IV: Performance of each action category on HHAR dataset.

GPT4 performs further analysis of the input raw data, which leads to the conclusion that the data is characterized by periodic and up-and-down movement. After accurate analysis and combining previous expert knowledge, in the fourth part, GPT4 makes an accurate judgment that the movement is most likely walking.

III. Special Properties of LLMs

Logical Reasoning Ability: Based on the findings of prior research, it has been established that LLMs possess the capacity for logical reasoning. Furthermore, our observations indicate that this ability extends beyond textual information processing and encompasses the analysis of data derived from the physical realm. Existing LLMs such as GPT4 are powerful adapters with the capability to convert unprocessed, raw digitized sensor data into abstract linguistic representations, such as periodic, stationary, abrupt, etc. This can be likened to an exceptionally potent filter that addresses the challenge of handling out-of-distribution [21] data in conventional machine learning or deep learning approaches. Rather than focusing on learning to match the specific characteristics of raw data, LLMs strive to generate the most plausible assertions by emulating human thought processes. However, different LLMs vary in their logical abilities. Fig. 6 shows the results of providing the same prompt to GPT3.5 [6], Google Gemini [7], and LLaMA2-70b [8]. We found that LLaMA2-70b is tentative to provide two possible answers after logical reasoning, instead of a deterministic answer.

Perfunctory Answer: There have been instances where ChatGPT has produced results that are considered perfunctory. In our tests, as shown in Fig. 4. We have observed that GPT4 occasionally provides constructive opinions without corresponding answers as well. It is crucial to mitigate this issue in order to achieve real-time and stable interaction between LLMs and CPS.

> Without using computational tools to visually plot or statistically analyze the data, the analysis must rely on recognizing patterns described above in the numerical data provided. Look for the characteristics in the given data that match the typical patterns of upstairs or downstairs movement, keeping in mind that real-world data can be noisy and may not perfectly fit theoretical expectations.

Figure 4: Perfunctory answer generated by GPT4.

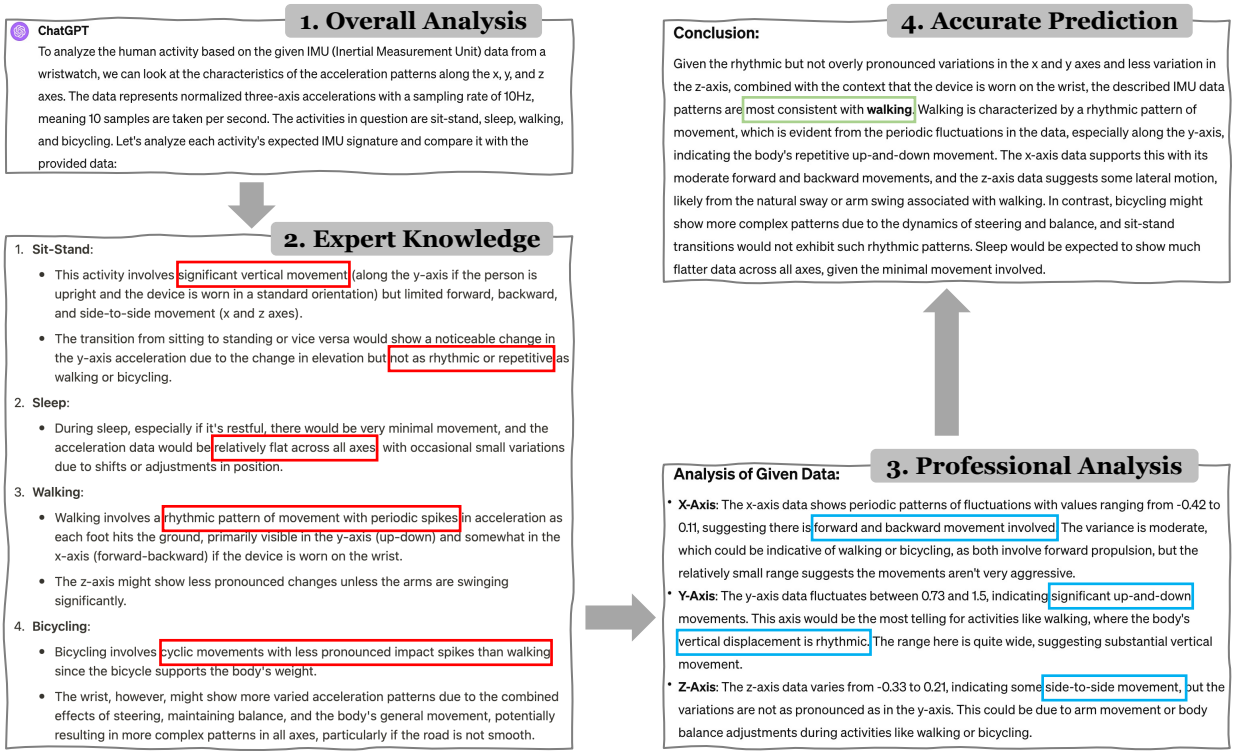
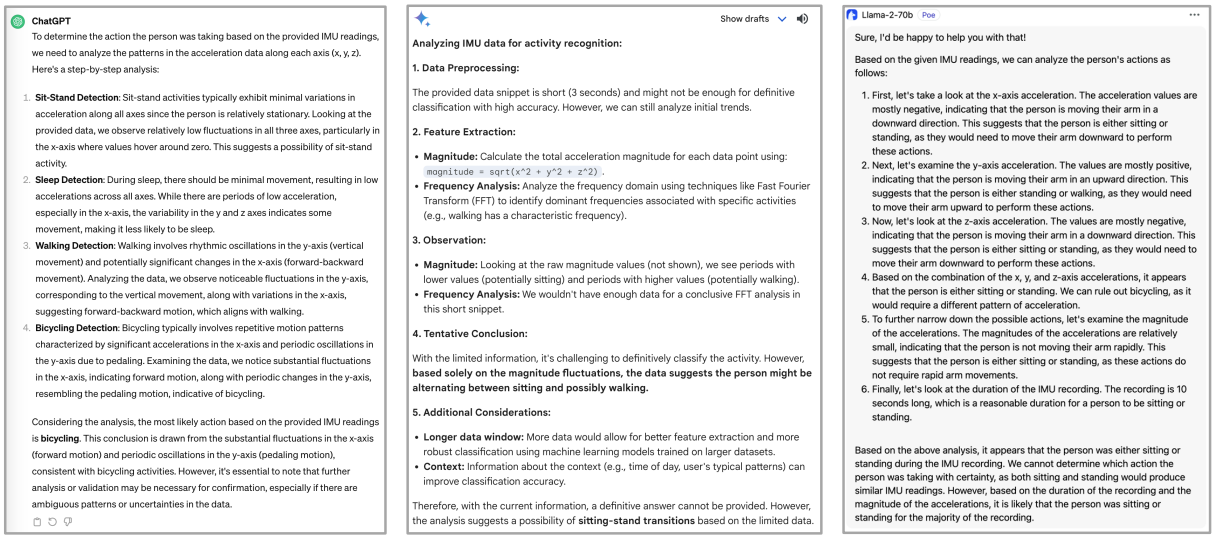


Figure 5: Detailed step-by-step inference generated by GPT4 with a walking example.



(a) GPT3.5 (b) Google Gemini (c) LLaMA2-70b

Figure 6: A comparison of the inference results generated by other LLMs for the walking scenario.

IV. Future Work and Conclusions

This study demonstrates that LLMs can be used as a foundational model to perform HAR in a zero-shot manner with high accuracy and robustness, which shows that LLMs are capable of processing IoT sensor data without the expertise-guided examples given. The findings highlight the capability of LLMs to analyze the raw sensor data and imply a transformative impact on the Cyber-Physical Systems domain. Nevertheless, further study should be carried out to assess to what extent and in what case the LLMs are effective. Standardizing the evaluation process with a more comprehensive set of evaluation procedures and benchmarks is needed. By deepening our comprehension

of the capabilities and constraints of Language Models, we can harness their prowess to advance our comprehension of real-world data, including abstract data like the WiFi Channel State Information.

ACKNOWLEDGMENT

This paper is supported in part by the NSFC under grant No. 62222216 and Hong Kong RGC ECS under grant 27204522.

REFERENCES

[1] T. Brooks, B. Peebles, C. Homes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[2] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, 2022.

[3] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun *et al.*, "Personal llm agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024.

[4] M. Xu, N. Dusit, J. Kang, and e. a. Xiong, Zehui, "When large language model agents meet 6g networks: Perception, grounding, and alignment," *arXiv preprint arXiv:2401.07764*, 2024.

[5] H. Xu, L. Han, Q. Yang, M. Li, and M. Srivastava, "Penetrative ai: Making llms comprehend the physical world," in *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, 2024, pp. 1–7.

[6] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[7] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[9] S. Chan Chang, R. Walmsley, J. Gershuny, T. Harms, E. Thomas, K. Milton, P. Kelly, C. Foster, A. Wong, N. Gray *et al.*, "Capture-24: Activity tracker dataset for human activity recognition," 2021.

[10] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *Proceedings of the 13th ACM conference on embedded networked sensor systems*, 2015, pp. 127–140.

[11] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *Ijcai*, vol. 15. Buenos Aires, Argentina, 2015, pp. 3995–4001.

[15] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen, "Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications," *GetMobile: Mobile Computing and Communications*, vol. 26, no. 3, pp. 39–42, 2022.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[17] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *arXiv preprint arXiv:2311.16452*, 2023.

[18] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[19] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," *arXiv preprint arXiv:2401.06866*, 2024.

[20] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[21] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.