Gaze2Prompt: Turning Eye-Tracking Data into Visual Prompts for Multimodal LLMs

Jae Young Choi
KAIST
Daejeon, Republic of Korea
jaeyoungchoi@kaist.ac.kr

Seon Gyeom Kim
KAIST
Daejeon, Republic of Korea
ksg_0320@kaist.ac.kr

Jaywoong Jeong
KAIST
Daejeon, Republic of Korea
jaywoong.jeong@kaist.ac.kr

Ryan Rossi
Adobe Research
San Jose, California, USA
ryrossi@adobe.com

Jihyung Kil
Adobe Research
San Jose, California, USA
jkil@adobe.com

Tak Yeon Lee
KAIST
Daejeon, Republic of Korea
reflect9@gmail.com

Abstract

Large Language Models (LLMs) are increasingly being adapted to interpret the physical world through sensor data. However, feeding raw sensor data (e.g., eye-tracking) into these models as text prompts can lead to excessive token overhead and degraded model performance. This study propose an alternative approach: transforming eye-tracking data into visual representations that serve as prompts for Multimodal Large Language Models (MLLMs). Specifically, we explore three types of visualizations - time-series plot, scanpath, and heatmap - and evaluate their effectiveness on a six-class eye-tracking classification task under zero-shot and one-shot conditions. Our results show that visual prompts not only reduce input tokens by over 85%, but also significantly improve accuracy. Especially heatmap achieved the highest one-shot accuracy of 73.9%, nearly doubling the 37.8% accuracy of raw text. These findings highlight the potential of visual abstraction for efficiently integrating trajectory-based sensor data into MLLM-powered process.

CCS Concepts

• **Human-centered computing** → *Ubiquitous and mobile computing*; *Visualization*; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Multimodal LLMs, Cyber-Physical Systems, Penetrative AI, Eye-Tracking Data Visualization

ACM Reference Format:

Jae Young Choi, Seon Gyeom Kim, Jaywoong Jeong, Ryan Rossi, Jihyung Kil, and Tak Yeon Lee. 2018. Gaze2Prompt: Turning Eye-Tracking Data into Visual Prompts for Multimodal LLMs. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

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

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The capabilities of Large Language Models (LLMs) are evolving to comprehend the real-world phenomena including human behavior [18]. Recent studies used sensor data collected via EEG, accelerometer, 3-D LiDAR, and ECG as input for LLMs [3, 7–9, 19–21]. The primary goal of these studies was to explore how LLMs can perform classification, reasoning, data processing, and prediction tasks using sensor data. For example, Xu et al. [18] demonstrated that LLMs can discern user motion and detect heartbeat peak without task-specific feature engineering. Based on this, Ji et al. [8] evaluated GPT-4 on a human activity recognition task in a zero-shot setting. Furthermore, Xue et al. [19] and Gruver et al. [7] showed that LLMs can process time-series data for forecasting tasks, highlighting their potential for sequence modeling.

Sensor data is commonly integrated into LLMs by representing it as raw text in input prompts [10]. However, this approach can lead to reduced model performance when processing long contexts, as well as high token consumption - resulting in high computational and financial costs [11, 22]. In response to these challenges, Yoon et al. [22] transformed sensor data into chart images and used them as visual prompts for Multimodal Large Language Models (MLLMs). This approach effectively condensed long data sequences into single image, reducing token overhead and enhancing the accuracy of classification tasks. In previous study, sensor data such as accelerometers, ECG, EMG, respiration, and audio have been visualized using raw signal plots or spectrograms as inputs for MLLMs [4, 22]. While these visualizations are tested on time-series data that represents signal amplitude over time or frequency components, spatially grounded data such as trajectory or tracking data have not been explored in this context. For instance, eye-tracking data contains two-dimensional spatial information in addition to temporal dynamics, which makes it fundamentally different from time-series data. This work explores the viability and implications of representing eye-tracking data as visual inputs for MLLMs.

In data visualization, eye-tracking data is often represented using scanpaths and heatmaps [1, 15]. Scanpaths illustrate the sequence of gaze movements as thin lines, conveying both spatial positions and temporal progression. In contrast, heatmaps provide a color-coded overview of attention distribution, emphasizing spatial density while omitting temporal information. Alongside these methods, this study also incorporates a time-series line plot, which directly depicts gaze coordinates against a time axis to highlight temporal

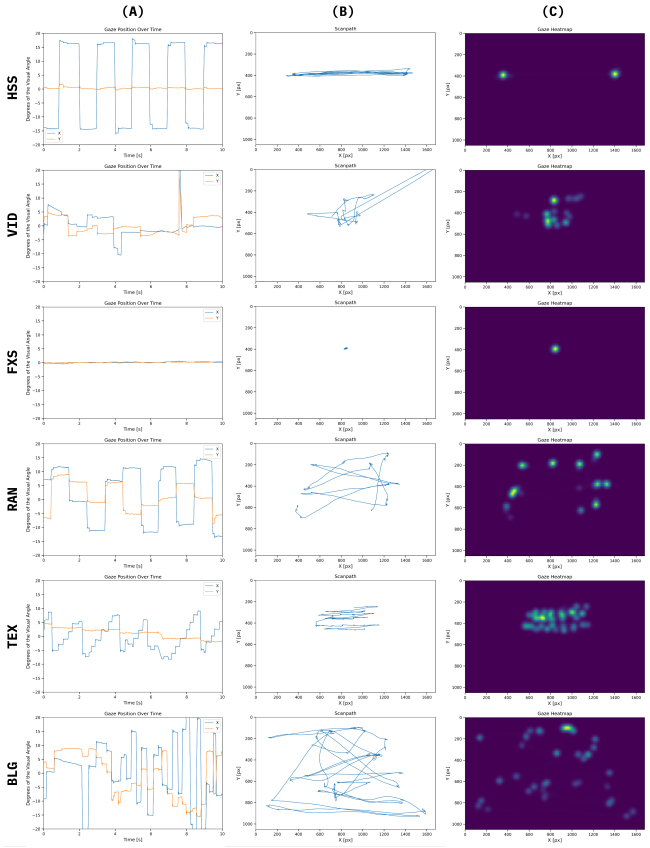

Figure 1: Three visualization types—(A) time-series plot, (B) scanpath, and (C) heatmap—are used to represent eye-tracking data across six activity categories (shown in rows).

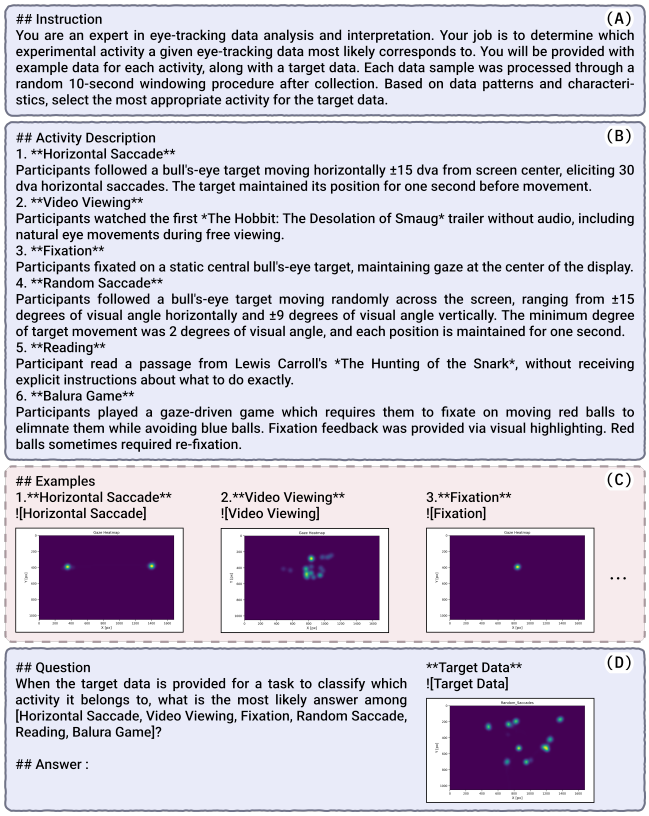

Figure 2: Prompt structure used in experiments. (A) Instruction prompt (B) Explanation of six categories (C) Example visual prompts (only for one-shot setting) (D) Question prompt with target data.

dynamics. Accordingly, this study investigates (1) whether visual prompts can serve as a more efficient and effective alternative to raw text, and (2) how the performance of different representations varies across behavioral categories, and what this reveals about MLLM's visual interpretation process.

Our results are as follows. First, visual prompts outperform text-based raw data in both accuracy and token efficiency. Second, the heatmap demonstrates strong overall performance, whereas the effectiveness of each visualization varies depending on the specific activity. It highlights the importance of choosing an appropriate visual encoding that matches the specific nature of the activity being analyzed, rather than relying on one-size-fits-all visualization. Our contribution extends the visual prompting paradigm to spatio-temporal eye-tracking data and offers a comprehensive analysis that reveals how the effectiveness of visualization methods varies by activity categories.

2 Methodology

2.1 Data Preparation

This study utilized *GazeBase* dataset [6], a comprehensive longitudinal dataset consisting of 12,334 monocular eye movement recordings from 322 participants. The eye-tracking data were recorded at 1 ms intervals, with gaze positions measured in degrees of visual angle (*dva*). Each subject demonstrated six distinct eye-tracking activities: Horizontal Saccade (*HSS*), Video Viewing (*VID*)[1], Fixation (*FXS*), Random Saccade (*RAN*), Reading (*TEX*), and Playing Balura Game (*BLG*). The dataset was collected longitudinally over nine rounds, with participant numbers decreasing from 322 to 14 in the final (9th) round. Based on our investigation, recordings from the later rounds tend to be much cleaner. We thus sampled two subsets—an *one-shot example* and a *test set*—for the classification experiment accordingly. In specific, the *one-shot example* consists of eye-tracking data of a single randomly selected subject from the final round. In contrast, the *test set* comprises 30 randomly-selected participants from the first round, to ensure a robust evaluation of the model's generalization. The selected participants are demographically balanced (15 females, 15 males; $M_{\text{age}} = 21.87$, $SD_{\text{age}} = 4.08$) and reflect the overall composition of the original dataset.

To maintain a consistent input length across samples, each recording was trimmed to a 10-second window, and then converted into four commonly used representations[2, 16]: raw text data, time-series plot, scanpath, and heatmap. All numeric values in the raw text data were rounded to two decimal places to maintain conciseness and avoid unnecessary precision, and the chart images were uniformly resized to 1229 × 768 pixels to ensure consistency.

[1]Although the original dataset contains two video viewing activities, we used the first one only, to ensure distinct activity categories and avoid redundancy.

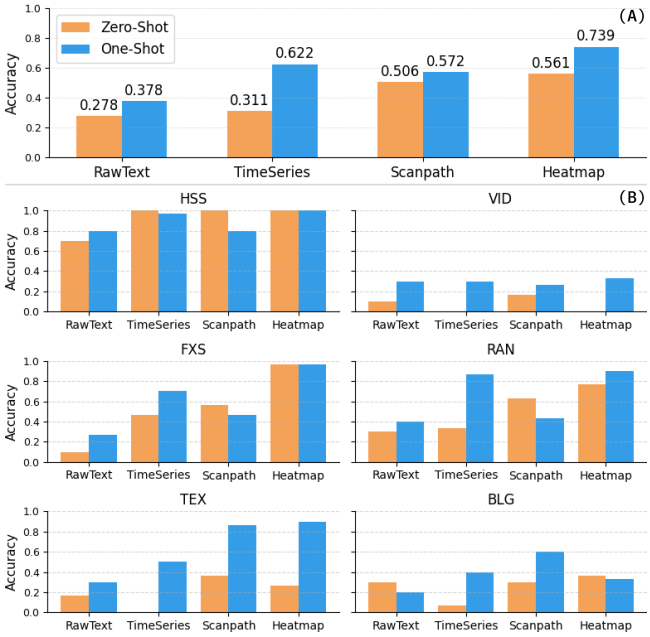

Figure 3: Accuracy comparison between zero-shot and one-shot settings across various data representations, showing (A) overall and (B) per-activity categories.

2.2 Experimental Design

To compare the effectiveness of the four data representations, we performed a standard classification experiment targeting six predefined eye-tracking activities. The experiments were carried out using GPT-4o-2024-11-20 [13]. To better understand the process of answering, we designed a structured output [14] using Chain-of-Thought method [17] to generate step-by-step reasoning. In order to examine the model's ability to generalize to new tasks and the performance gain achievable with minimal supervision, we created both zero-shot and one-shot prompts. The prompts consists of four sections, as shown in Figure 2. First, the *Instruction* section provides the model with context and guidance, followed by *Activity Description* that outlines the characteristics of the six classes of eye-tracking activity. Only the one-shot prompt has the *Examples* section that provides six example representations. Lastly, both prompts conclude with the *Question* section that presents the target instance to be classified. To maintain consistency between modalities, both the visual and raw text inputs were uniformly labeled as "data". Furthermore, to mitigate ordering bias [23], we rotated the order of descriptions and examples for each activity.

3 Results

Text vs Visual Representations. As shown in Figure 3(A), all visual representations—time series plot, scanpath, and heatmap—consistently outperformed the raw text input, except for *VID* activity where no representations achieved over 0.4 accuracy. In general, the heatmap achieved the highest accuracy in both zero-shot and one-shot settings, reaching 0.561 and 0.739 respectively. However, *BLG* was an exceptional case where one-shot scanpath achieved the higher accuracy as shown in Figure 3(B).

Zero-shot vs One-shot. The one-shot prompt led to higher accuracies in general. For instance, the time series plot, which achieved 0.311 in the zero-shot setting, doubled its accuracy to 0.622 in the one-shot condition. However, for scanpath visualizations, one-shot prompting did not outperform zero-shot for three activities (*HSS*, *FXS*, and *RAN*), indicating that the added example did not reliably enhance classification accuracy.

Activity Categories. An analysis of performance by behavioral category, as detailed in Figure 3(B), reveals several key patterns. Overall, the heatmap representation yielded the highest accuracy for tasks such as *HSS*, *FXS*, and *RAN* in both settings. While one-shot learning generally improved accuracy, an exception was observed for *HSS* task using time-series plot and scanpath, *FXS* and *RAN* tasks using scanpath, and *BLG* task using raw text and heatmap. Conversely, the *VID* task seemed to be challenging in all circumstances. For the *BLG* task in the one-shot setting, the accuracies of scanpath (0.60; 18 of the 30 cases) and time-series plot (0.40) were higher than the accuracy of heatmap (0.33).

The confusion matrices in Figure 4 illustrate specific misclassification patterns. In the zero-shot condition, the time-series plot commonly misclassified other activities as *HSS*; for instance, 27 *TEX* and 20 *RAN* cases. For one-shot condition, a common error across all visual methods was incorrectly classifying *VID* as *BLG* cases, which occurred 13 times for time-series, 13 for scanpath, and 10 for heatmap (N=30 for each activity). A mutual misclassification pattern was also observed between *BLG* and *RAN*. *BLG* was often mislabeled as *RAN*, a tendency prominent in the zero-shot setting (e.g., 18 out of 30 for both time-series plot and heatmap) but still present in the one-shot condition (e.g., 15 out of 30 for time-series plot and 19 out of 30 for heatmap). Conversely, *RAN* was also frequently misclassified as *BLG*, especially in the one-shot scanpath results (9 out of 30 cases).

Input Token Efficiency. Token counts for text inputs were calculated by using `cl100k_base` encoding scheme [12]. Each visualization image was automatically splitted into six 512×512 tiles, with each image using 1,105 tokens[2]. In the result, the use of visual representations was confirmed to be substantially more efficient in token consumption than using raw text input. In the zero-shot condition, raw text input required an average of 10,148 tokens, whereas all three visual prompts consistently used only 1,527 tokens. Similarly, in the one-shot condition, raw text input consumed an average of 69,928 tokens, while the visual prompts required 8,294 tokens. By using visual prompts, token consumption in this experiment was reduced to 15.05% and 11.96% of the raw text in the zero-shot and one-shot settings, respectively.

4 Discussions

Efficacy of Turning Eye-Tracking Data into Visual Prompts. Our findings confirm that for eye-tracking data, visual prompting is an effective strategy for enhancing classification accuracy while substantially reducing token usage, extending prior work into the spatio-temporal domain of gaze analysis. We evaluated multiple visualization techniques and discovered that the heatmap representation consistently yielded the highest accuracy. This result is noteworthy because heatmaps lack explicit temporal information. This suggests that for gaze-based tasks, a concise summary of spatial attention can be a powerful feature for an MLLM than a verbose sequence of raw gaze coordinates.

[2]Based on GPT-4o's calculation: 85 base tokens + 170 tokens per tile.

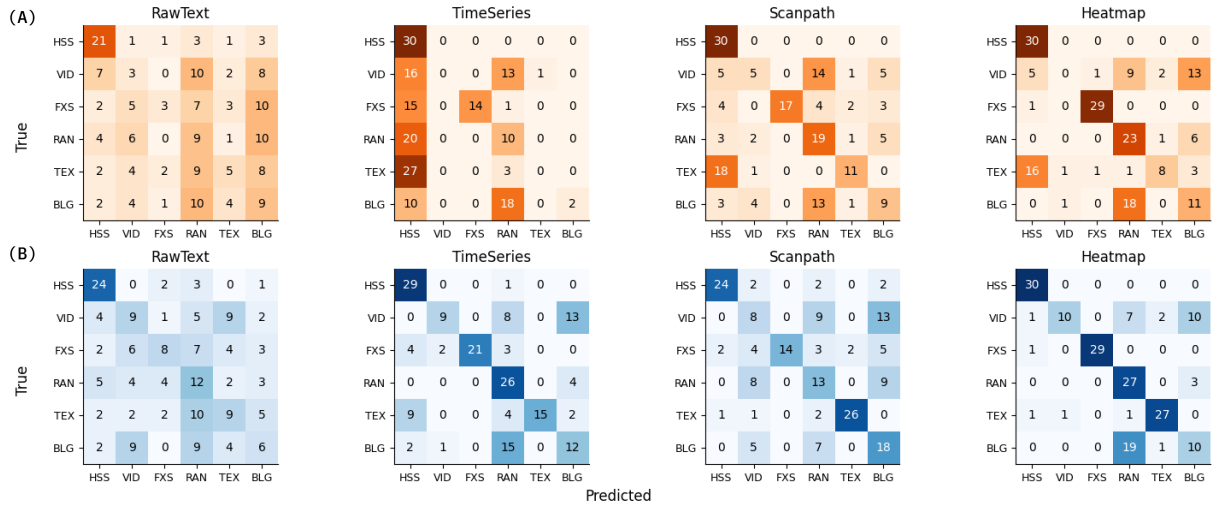

Figure 4: Confusion matrices of four data representation types in (A) zero-shot setting and (B) one-shot setting.

Aligning Visualizations with Behavioral Characteristics. A more detailed observation is that the optimal visualization method varies with the nature of the behavioral categories. For instance, in activities that create spatially predictable and dense fixation zones, such as the point-to-point movements in *HSS*, *RAN*, and *TEX*, or the single focus point in *FXS*, the heatmap's ability to summarize these high-density areas proved superior.

However, for *BLG*, both time series plot and scanpath representations achieved higher accuracy than the heatmap in one-shot condition. This activity involves tracking dynamically moving objects, a process where temporal and sequential information is critical. The MLLM's Chain-of-Thought reasoning supported this, noting that one-shot scanpath revealed *"Observed the pattern of dense and intricate movements without a strongly defined pattern."*. In contrast, all visual methods performed poorly on *VID*, frequently misclassifying it as *RAN* and *BLG*. One of the reasoning steps for zero-shot time-series states: *"From the image of the eye-tracking data, we observe multiple saccades with significant movement in both X and Y axes."*. The free viewing nature of *VID* lacks a distinct gaze signature, making it difficult to classify without additional context while the model can capture the shared high-level feature of dynamic movement presented both in *VID* and *BLG*. This suggests that for less structured activities, gaze data alone may be insufficient for MLLMs to distinguish the subtle differences.

5 Limitation and Future Work

Enhancing Visual Representations. The visualizations used in this study were foundational. The scanpath, for instance, illustrates the connections between gaze points but do not explicitly represent the temporal order of movements or the duration of fixations. Techniques such as varying marker sizes [1] to indicate dwell time were not incorporated, limiting the information available on fixation points. Similarly, our heatmaps were generated without being overlaid on a specific stimulus, which is a common practice in user interface analysis [5]. However, given the strong performance of the informationally simple heatmap, this future work should also systematically investigate the trade-off between information density and abstractive clarity. Determining the optimal level of detail that aids MLLM reasoning remains a topic for future exploration.

Expanding the Scope of Eye Tracking Tasks. This study focused on a activity classification task based on gaze data, which is not the only problem addressed with eye-tracking. There is an opportunity to investigate whether visual prompting can be effective for other common eye-tracking analyses. Future research could explore the application of this method to tasks such as saliency prediction, user identification, or cognitive load estimation from gaze patterns. Moreover, as MLLMs can interpret natural language instructions, their integration into cyber-physical systems could facilitate more open-ended task execution and complex reasoning, unconstrained by predefined analytical scopes.

Generalization to Other Trajectory Data. The methodology of representing tracking data as visual inputs for MLLMs holds promise beyond eye-tracking. This approach can be generalized to a wide variety of trajectory tracking data. For example, a user's physical movement traces from wearable sensors, the navigation paths of robots or drones, or GPS based vehicle logs could be similarly visualized. Such data could be rendered as two dimensional or three dimensional path plots, which are spatially grounded. This could unlock advanced capabilities such as path forecasting, anomaly detection, or generating explanations for observed movements. Ultimately, this approach could significantly advance how MLLMs process and reason about dynamic events from sensor data, offering a new method for analyzing real-world's physical behaviors.

6 Conclusion

This study investigated the effectiveness of using image representations of eye-tracking data as visual prompts for MLLMs. Our experiments demonstrated that visualization methods, particularly heatmaps, significantly outperform raw text inputs in a classification job setting, achieving higher accuracy while dramatically reducing token consumption. This work confirms that transforming complex sensor data into an appropriate visual format is a viable and efficient strategy for enabling MLLMs to interpret nuanced human behaviors from trajectory data. Furthermore, our finding that different visualizations excel at different tasks suggests that future research should focus on tailoring visual representations to specific analytical goals to fully leverage the capabilities of reasoning and understanding real world context of MLLMs.

References

[1] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2014. State-of-the-art of visualization for eye tracking data.. In *Eurovis (stars)*. 29.

[2] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2017. Visualization of eye tracking data: A taxonomy and survey. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 260–284.

[3] Simon Böhi and Shkurta Gashi. 2024. Large Language Models for Wearable Data Analysis and Interpretation. In *Tiny Paper ICLR 2024*.

[4] Satvik Dixit, Laurie Heller, and Chris Donahue. 2024. Vision Language Models Are Few-Shot Audio Spectrogram Classifiers. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

[5] Vanessa Georges, François Courtemanche, Sylvain Senecal, Thierry Baccino, Marc Fredette, and Pierre-Majorique Leger. 2016. UX heatmaps: mapping user experience on visual interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4850–4860.

[6] Henry Griffith, Dillon Lohr, Evgeny Abdulin, and Oleg Komogortsev. 2021. GazeBase, a large-scale, multi-stimulus, longitudinal eye movement dataset. *Scientific Data* 8, 1 (2021), 184.

[7] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2023), 19622–19635.

[8] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. Hargpt: Are llms zero-shot human activity recognizers?. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 38–43.

[9] 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. In *Proceedings of the fifth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 248)*. PMLR, 522–539.

[10] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).

[11] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

[12] OpenAI. 2022. tiktoken. https://github.com/openai/tiktoken.

[13] OpenAI. 2024. *GPT-4o (Version 2024-11-20)*. https://openai.com

[14] OpenAI. 2024. Introducing Structured Outputs in the API. https://openai.com/index/introducing-structured-outputs-in-the-api/.

[15] Michael Raschke, Tanja Blascheck, and Michael Burch. 2013. Visual analysis of eye tracking data. In *Handbook of human centric visualization*. Springer, 391–409.

[16] Lisa-Marie Vortmann, Jannes Knychalla, Sonja Annerer-Walcher, Mathias Benedek, and Felix Putze. 2021. Imaging Time Series of Eye Tracking Data to Classify Attentional States. *Frontiers in Neuroscience* Volume 15 - 2021 (2021). doi:10.3389/fnins.2021.664490

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[18] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.

[19] Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (2023), 6851–6864.

[20] Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Suyuan Zhao, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Ping Zhang, et al. 2025. ECG-LM: Understanding Electrocardiogram with a Large Language Model. *Health Data Science* 5 (2025), 0221.

[21] Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. 2025. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9247–9255.

[22] Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2219–2241.

[23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009