Penetrative AI: Making LLMs Comprehend the Physical World

Huatao Xu[†], Liying Han[§], Mo Li[*†], Mani Srivastava[§]
[*]Hong Kong University of Science and Technology,
[†]Nanyang Technological University, [§]University of California Los Angeles
Email:huatao001@ntu.edu.sg, {liying98, mbs}@ucla.edu, lim@cse.ust.hk

ABSTRACT

Recent developments in Large Language Models (LLMs) have demonstrated their remarkable capabilities across a range of tasks. Questions, however, persist about the nature of LLMs and their potential to integrate common-sense human knowledge when performing tasks involving information about the real physical world. This paper delves into these questions by exploring how LLMs can be extended to interact with and reason about the physical world through IoT sensors and actuators, a concept that we term "*Penetrative AI*". The paper explores such an extension at two levels of LLMs' ability to penetrate into the physical world via the processing of sensory signals. Our preliminary findings indicate that LLMs, with ChatGPT being the representative example in our exploration, have considerable and unique proficiency in employing the knowledge they learned during training for interpreting IoT sensor data and reasoning over them about tasks in the physical realm. Not only this opens up new applications for LLMs beyond traditional text-based tasks, but also enables new ways of incorporating human knowledge in cyber-physical systems.

1 INTRODUCTION

Large Language Models (LLMs) have made remarkable strides [1, 15, 20]. A particularly revolutionary milestone in this domain is ChatGPT [12], which excels in fluid, human-like conversations, marking a new era in human-AI interactions. These latest LLMs cultivated on extensive text datasets have showcased remarkable capabilities across diverse tasks, including coding and logical problem-solving [2].

However, beneath these impressive feats lie an intriguing question: Are LLMs the result of mere statistical memorization, or do they already organically assimilate common-sense human knowledge in a pervading "world model[i]"? Does such a world model, if existing, enable the LLMs to establish sophisticated connections among high-level concepts and unravel the intricacies of the world's operations?

[i]We adopt the "world model" definition in [7] which is referred to as an internal model of how the world works, comprising enormous amounts of common sense knowledge.

Mo Li and Mani Srivastava are the corresponding authors.

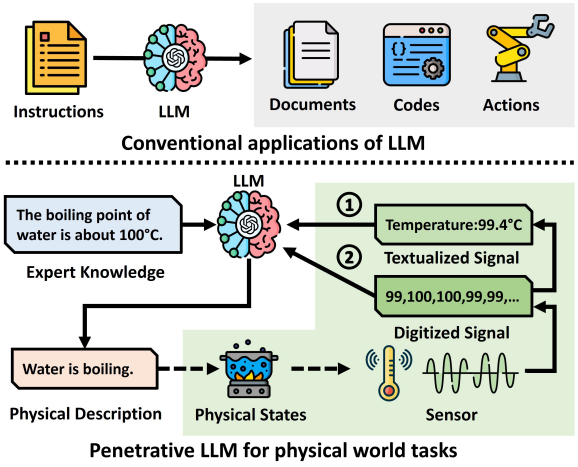


Figure 1: Overview of Penetrative AI.

In this paper, we hypothesize that the answer to the above question is true, and extend the boundaries of LLM capabilities by directly letting them interact with the physical world through Internet of Things (IoT) sensors. A basic example of this process is depicted in Figure 1, where different from the conventional way of LLM usage in natural language tasks, an LLM is expected to analyze sensor data and aided by expert knowledge to deduce the physical states of the object. These sensor and actuator readings are indeed projections from the physical world, and the LLM is expected to harness its common human knowledge to comprehend sensor data and execute perception tasks.

As illustrated in Figure 1, we formulate such a problem from a signal processing's point of view, and specifically explore the LLMs' penetration into the physical world at two signal processing levels of the sensor data: i) with the textualized signal derived from underlying sensor data, and ii) with the digitized signal, essentially numerical sequences of raw sensor data. We term this endeavor "***Penetrative AI***" – where the LLM-based "world model" transforms into sentience, seamlessly integrating with the Cyber-Physical Systems (CPS) to access and interact with the physical world.

Our methodology is exemplified through two illustrative applications at two different levels, respectively - user activity sensing where textualized signals from smartphone

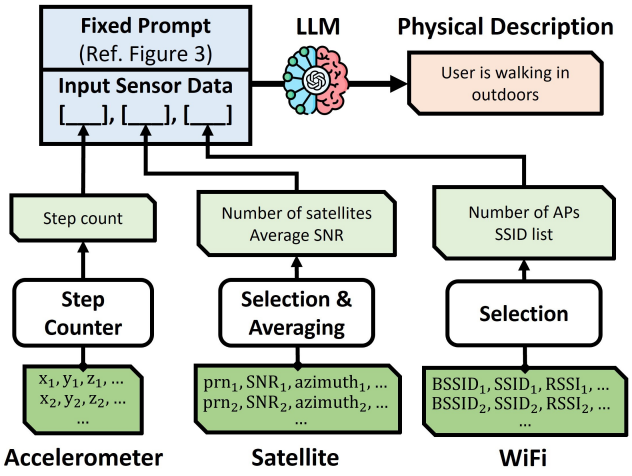


Figure 2: Overview of user activity sensing with LLM

accelerometer, satellite, and WiFi data are analyzed to discern user motion and environment conditions, and human heartbeat detection where digitized electrocardiogram (ECG) data are utilized to derive the heartbeat rate. Preliminary findings are encouraging, showcasing LLMs' proficiency in interpreting IoT sensor data and performing perception tasks in the physical world. Our exploration also underscores that existing LLMs, such as ChatGPT-4, may already possess the capability to establish intricate connections among world knowledge and can be guided to tackle CPS tasks.

Section 2 and Section 3 will elaborate on the design and experiment results of these two illustrative applications. Following that, in Section 4 we share our insights on the potentials of penetrative AI, propose a general technical framework, and discuss foreseeable challenges to advance this burgeoning research frontier. We present related works in Section 5 and conclude this paper in Section 6.

2 PENETRATIVE LLM WITH TEXTUALIZED SIGNALS

This section describes our effort in tasking ChatGPT, a chosen vehicle, to comprehend IoT sensor data at the textualized signal level.

2.1 An Illustrative Example

We take activity sensing as an illustrative example, where we task ChatGPT with the interpretation of sensor data collected from smartphones to derive user activities. The input sensor data encompass smartphone accelerometer, satellite, and WiFi signals, and the desired output is to discern the user motion and environment context. Figure 2 presents the overview of this LLM-based design – the sensor data are

Objective:
Determine a user's motion and surrounding conditions by ...

Sensor Data and Expert Knowledge:
Here's how to interpret various sensors data:
1. A high step count means ...
2. A high satellite count and carrier-to-noise density indicates ...
3. A large count of APs suggests ...., SSIDs can hint at specific locations...

Reasoning Example:
Sensor data:
1. Step count: 5/min.
2. Satellites detected: 0. Average carrier-to-noise: 0.0dB.
3. Total WiFi APs scanned: 10. SSID list: ['McDonald's Singapore', ...].
Reasoning:
1. The low step count indicates the user is stationary.
2. No detected satellites and low carrier-to-noise suggest an indoor environment.
3. Detected WiFi APs point to an indoor setting, with SSIDs like 'McDonald's Singapore' hinting at a location close to McDonald in Singapore.
Summary: The user is stationary, likely inside a McDonald restaurant in Singapore.

Now infer a user's motion and surrounding conditions with the following sensor data: ...

Figure 3: A brief prompt for activity sensing.

pre-processed by individual sensing components and the textualized sensor states are supplied to the LLM with a fixed prompt for activity inference.

Objective and rationale. We convey a clear task to ChatGPT – "determine a user's motion and surrounding conditions by analyzing sensor data from their smartphones". The basic idea is that when the user conducts different activities in different environments, the collected sensor data would exhibit varied patterns, which can be analyzed to infer users' activities.

Data preparation. To facilitate ChatGPT comprehension of the sensor data, we undertake a preprocessing step where raw data from different sensing modules are separately converted into textualized states that are expected interpretable by ChatGPT. Figure 2 illustrates such a step.

We employ an existing step counting algorithm [13] to pre-process accelerometer readings (6,000 samples from 10 seconds of XYZ acceleration sampled at 200Hz) and derive the step count – which transforms the 6,000 raw data points into a single state, which can be textually expressed, e.g., "step count: 5/min".

The Android operating system offers a comprehensive set of Global Navigation Satellite System (GNSS) satellite measurements [3], including information like Pseudo-Random Noise (PRN) as a satellite identifier, constellation type, Signal-to-Noise Ratio (SNR), and many others. To streamline the data for ChatGPT interpretation, we filter and distill the satellite data into two key attributes: the number of detected satellites and their average SNR.

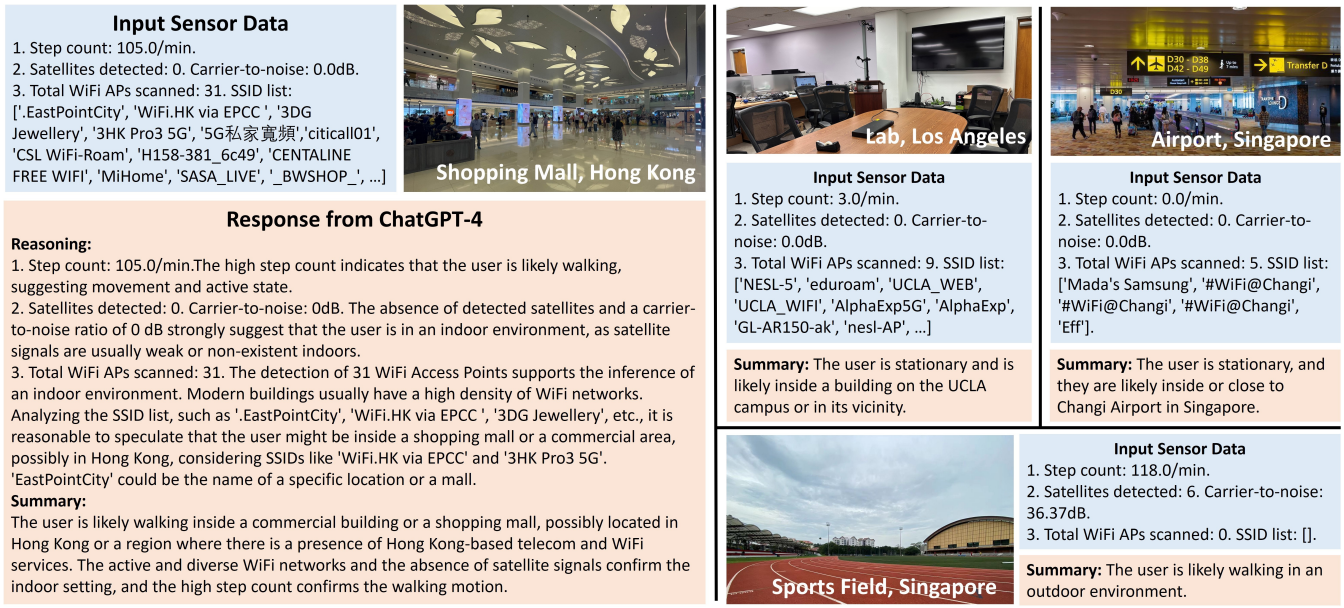


Figure 4: Response examples of ChatGPT-4 for activity sensing[ii].

The Android system supports scanning for nearby APs and provides comprehensive information about scanned APs [4]. Similar to satellite data, we disregard less relevant details and focus on critical information – Service Set Identifier (SSID) and Received Signal Strength Indicator (RSSI). To streamline the data and reduce text length, we further filter APs with an RSSI lower than -70 and instruct ChatGPT to analyze the SSIDs to capture useful location information.

Expert knowledge. In addition, we provide guidance to LLMs by including explicit text-based descriptions of the relationships between sensor patterns and user activity states in the prompts, as illustrated in Figure 3. For instance, a high satellite count and carrier-to-noise density indicate an outdoor setting with strong satellite signals.

Reasoning examples. To enhance the proficiency of ChatGPT in completing the sensing task, we provide a set of reasoning examples. Each example includes the data for processing, a step-by-step reasoning process, and a brief summary of the ground truth context. Figure 3 illustrates this with the reasoning example section.

Complete prompt. As demonstrated in Figure 3, a full prompt includes a defined objective and expert knowledge of the sensor data, all in natural language. Essentially, the way we construct the prompt serves as a means to educate and instruct ChatGPT to interpret sensor data and formulate its answers into a concise format. We thereafter present the prompt with succinct textualized sensor data of novel queries to ChatGPT as shown in Figure 2, which we expect to generate the inference results as a concise description of the user's activity. Note that the prompt, once completed, is frozen and we simply supply new textualized sensor data for new inferences without altering the prompt any further.

2.2 Preliminary Experiment Results

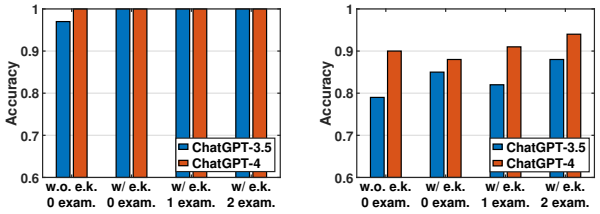
We conducted preliminary experiments in various scenarios – on university campuses, commercial buildings, subway stations, outdoor spaces, and across cities. All sensor data were collected using a Samsung Galaxy S8 Android smartphone. Accelerometer data was sampled at 100 Hz, while the satellite and WiFi data were sampled at 0.2 Hz. To perform our analysis, we utilized sensor data gathered from time windows spanning durations of 10 to 45 seconds and selected the most recent satellite and WiFi scanning results. The evaluation was carried out using both ChatGPT- 3.5 and ChatGPT-4 [12], accessible through the OpenAI API [11] and default parameter settings.

Figure 4 provides several example answers together with their ground-truth contexts. Due to space limits, we have omitted the detailed reasoning of ChatGPT responses except for the first example.

To quantitatively assess the efficacy of this approach, we tasked ChatGPT to explicitly provide the states of motion (between "stationary" and "motion") and environment (between "indoors" and "outdoors"). We experiment with varied settings – with/without expert knowledge in the prompt, as well as with different numbers of reasoning examples.

Figure 5(a) summarizes the accuracy for motion detection, which suggests two models perform reasonably well.

[ii]Due to the space limit, check more examples and the complete prompt at https://dapowan.github.io/wands_penetrative-ai/



(a) Motion detection. (b) Indoor/Outdoor.

Figure 5: Recognition accuracy of ChatGPT in activity sensing. The 'e.k.' denotes expert knowledge, and 'exam.' refers to the reasoning examples provided.

ChatGPT-3.5 occasionally outputs 'unknown' states leading to 97% accuracy even under the output constraints, which can be improved to 100% when detailed expert knowledge and reasoning examples are used in the prompt.

Figure 5(b) summarizes the accuracy for environment classification, which depends on multimodal sensor data fusion and therefore appears more challenging. Nevertheless, improved performance is observed when expert knowledge and more reasoning examples are used in the prompt. ChatGPT-4 achieves above 90% accuracy with the best prompt template.

Overall, the above experiment results suggest LLMs are highly effective in analyzing physical world signals when they are properly abstracted into textual representations. These findings align with our initial expectations based on the "world model" hypothesis of LLMs.

3 PENETRATIVE LLM WITH DIGITIZED SIGNALS

This section describes our effort to go beyond the general expectations of the textualized signal processing ability of LLMs. We specifically study the potential of ChatGPT in comprehending digitized sensor signals.

3.1 An Illustrative Example

We take human heartbeat detection as an illustrative example, where we task ChatGPT with the input of ECG waveforms to derive the heartbeat rate. Fundamentally different from the previous example, all sensor data in this application are expressed as sequences of digitized samples. Figure 6 provides an overview of the design.

Objective and rationale. The sensor data consist of a numerical sequence representing an ECG waveform. Our objective for ChatGPT is to identify the "R-peaks" [19], which are tall upward deflections in ECG data and correspond to the red dots in Figure 6. The objective part of the prompt succinctly states: "Find the R-peaks in an ECG waveform". An interesting and challenging job in this application is, we incorporate expert knowledge directly into the prompts, delegating the signal processing task to ChatGPT.

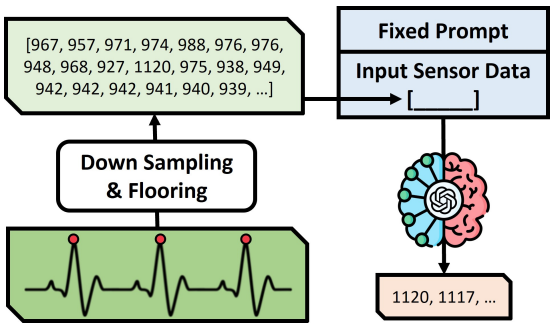


Figure 6: Overview of heartbeat detection with LLM.

Data preparation. The original ECG data are collected at a high sampling rate, e.g., 360Hz. In our design, we downsample to 36 Hz and keep the integer part of raw ECG readings to reduce sequence length and the number complexity.

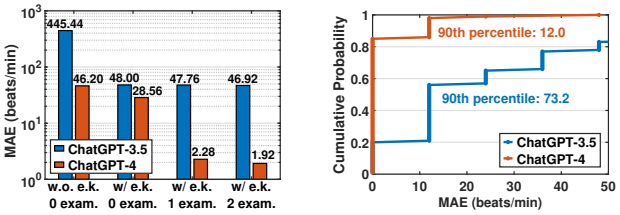
Expert knowledge. We provide interpretations regarding R-peaks: "An R-peak within a sequence of ECG numbers refers to a pronounced upward deflection, typically representing the largest and most conspicuous values within the sequence". We provide a natural language-based "algorithm" that LLMs understand to guide the selection of R-peaks. Three steps are included: i) assessing the overall range of ECG numbers, ii) identifying subsequences characterized by an initial lower value, a subsequent significant increase, and a return to the overall range, and iii) selecting the highest value from each such subsequence as the R-peak. We investigate whether ChatGPT can effectively execute such fuzzy logic (without explicit threshold values) when processing the digitized signals.

Reasoning examples. We also furnish ChatGPT with illustrative examples, encompassing digitized ECG data, a reasoning procedure, and a summary of R-peak numbers.

3.2 Preliminary Experiment Results

We conducted preliminary experiments with the MIT-BIH Arrhythmia Database [6], which is an ECG dataset equipped with ground truth annotations for R-peaks. We downsampled the raw ECG signal to 36 Hz and each input ECG data are from a 5-second window comprising 180 numbers. The evaluation was again carried out using both ChatGPT-3.5 and ChatGPT-4 with default parameters.

We use the Mean Absolute Error (MAE) to measure the deviation in beats per minute between the predicted and actual R-peaks. In Figure 7(a), we present our preliminary results of the two models under different settings. ChatGPT-4 consistently outperforms ChatGPT-3.5, achieving an MAE of 1.92 in the best case. We also examine the MAE of two models with knowledge and one reasoning example and



(a) Overall performance. (b) CDF of MAE.

Figure 7: MAE of ChatGPT in heartbeat detection. (a) The 'e.k.' denotes expert knowledge, and 'exam.' refers to the reasoning examples provided.

visualize the Cumulative Distribution Function (CDF) in Figure 7(b). The efficacy of ChatGPT-3.5 often generates prolonged sequences of R-peaks, resulting in substantial errors. Conversely, ChatGPT-4 demonstrates enhanced stability and precision in identifying R-peaks in the majority of cases.

In conclusion, our initial findings indicate that LLMs, particularly ChatGPT-4, exhibit remarkable proficiency in analyzing physical digitized signals when provided with proper guidance. These results provide additional evidence supporting the "world model" hypothesis of LLMs.

4 PENETRATIVE AI

While not achieving perfect accuracy, LLMs exhibit surprisingly encouraging performance, even when dealing with purely digital signals acquired from the physical world. This presents an enticing opportunity to leverage LLMs' "world model" and integrate it with IoT sensing capabilities to build intelligence into cyber-physical interactions—a concept we term as "Penetrative AI".

4.1 Penetrative AI Potentials

As the above example applications demonstrate, leveraging LLMs in a penetrative way may offer the following potentials.

Simplified solution deployment. This method allows users to interact with machines using plain language. It empowers effortless application design, eliminating the need for extensive programming expertise.

Enhanced data efficiency. Embedded with comprehensive knowledge, LLMs can achieve satisfactory results in new tasks with minimal examples, significantly reducing the demand for extensive data gathering and model training.

Explainable reasoning. With their natural language capabilities, LLMs enhance interpretability and provide transparent reasoning, empowering users to gain a deeper understanding of model decisions.

Handling fuzzy logic. LLMs can draw inferences from unclear and disorganized information similarly to humans, bypassing the strict requirement for precise logic.

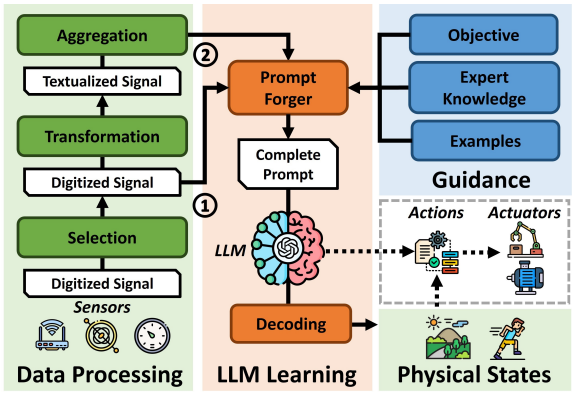


Figure 8: Framework overview.

Innovative multimodal fusion. By converting multimodal data into a scalable text-based format, our approach enables effortless adaptation to various tasks and modalities without the necessity for model reengineering

4.2 A Framework

Figure 8 depicts a structured framework for designing penetrative LLMs, which includes the following procedures: i) **Data processing** enables LLMs to interpret sensor data readily. Data transformation and aggregation may (or not) take place in textualizing (or simply digitizing) raw signal for LLM consumption. ii) **Guidance** supplies objectives, expert knowledge, or reasoning examples in the form of text. They help enhance LLM effectiveness in specific applications. iii) **LLM learning** forms comprehensive prompts, which are then utilized by LLMs to produce physical descriptions of entities or objects. In the dotted box of Figure 8, we also illustrate the potential for LLMs to intervene in the physical world using actuators. Refer to Section 4.3 for further details.

4.3 Challenges and Future Directions

Expanding LLM roles. Our current data processing relies on many human efforts. We can explore enhancing LLMs' roles by using their coding skills to generate symbolic programs for explicit signal processing.

From perception to action. While LLMs have shown promise in perception tasks, the next challenge is to assess their capability to actively intervene in CPS.

Enhanced prompt schemes. Our current approach employs independent prompts. However, prompts for complex tasks may need to interconnect and record states, enabling them to capture temporal and spatial relations in sensor data.

Tailored large models. We currently use generic LLMs without customization. A crucial future direction is to develop large models explicitly designed for IoT applications.

Augmenting numeric representation. Handling lengthy numeric sequences with LLMs can be inefficient, but this can

be enhanced by integrating representations from pretrained sensing models.

Natural language algorithm. LLMs enable us to design algorithms in natural language with fuzzy logic, as demonstrated in Section 3.1.3. Developing customized and highly effective natural language algorithms for LLMs represents another promising future avenue.

5 RELATED WORK

The significant strides in natural language processing show that large language models (LLMs) pretrained on vast datasets exhibit out-of-the-box capabilities [1, 15, 20]. Some works [5, 14] extend LLMs into multimodal models, e.g., vision-language-action models, for various tasks. Most such approaches often necessitate additional training, making them dependent on massive data collection for training and fine-tuning models to suit specific tasks.

A promising avenue In-Context Learning (ICL) leverages the inherent capabilities of LLMs to tackle new tasks by learning directly from input prompts without additional training [10]. Such a breakthrough opens up new applications for LLMs, e.g., Liu et al.'s work [9] in analyzing medical data for health-related tasks. Such applications typically rely on question-answer pair patterns for learning. In various domains such as gaming and robotics, LLM-based agents have harnessed the embedded general knowledge to generate actions or policies [8, 16–18]. These applications, however, primarily build on LLMs' language processing ability. Some also interface with the physical world, e.g., by robots [16], but in an indirect way and often through perception and control APIs defined in formal languages (programming codes).

In contrast to all existing efforts, this paper defines "*Penetrative AI*" which builds on the world model hypothesis of LLMs and treats them as sentient entities in comprehending physical phenomena and completing real-world tasks. In this undertaking, LLMs directly engage in perception tasks by processing IoT sensor signals at various levels. We believe this is the first effort to explore the boundaries of LLMs' ability to interact with the real physical world.

6 CONCLUSION

We present penetrative AI and explore the potential of leveraging large language models as world models to accomplish real-world tasks. Our findings illuminate a promising path for the integration of artificial intelligence and CPS, offering insights into the future of AI-powered solutions.

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020).

[2] Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712* (2022).

[3] Android Developers. 2023. GnssStatus. https://developer.android.com/reference/android/location/GnssStatus

[4] Android Developers. 2023. ScanResult. https://developer.android.com/reference/android/net/wifi/ScanResult

[5] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).

[6] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.

[7] Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* 62 (2022).

[8] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9493–9500.

[9] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. *arXiv preprint arXiv:2305.15525* (2023).

[10] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943* (2021).

[11] OpenAI. 2023. GPT. https://platform.openai.com/docs/guides/gpt

[12] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[13] Meng-Shiuan Pan and Hsueh-Wei Lin. 2014. A step counting algorithm for smartphone users: Design and implementation. *IEEE Sensors Journal* 15, 4 (2014), 2296–2305.

[14] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824* (2023).

[15] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[16] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11523–11530.

[17] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

[18] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[19] Frank G Yanowitz. 2010. Lesson III. Characteristics of the Normal ECG. *University of Utah School of Medicine* (2010).

[20] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).