Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Application of Large Language Models in Embodied Artificial Intelligence

Zitian Li*

School of Statistics and Mathematics, Shandong University of Finance and Economics, Shandong, China

* Corresponding Author: lizitian@berkeley.edu

Abstract. The convergence of Artificial Intelligence (AI) and robotics has led to the emergence of embodied AI, where intelligent systems equipped with sensors and actuators interact with the physical world and operate alongside humans. These systems are transforming industries such as autonomous driving, healthcare, and household assistance. However, despite extensive research, embodied AI systems face significant limitations, including poor generalization and performance degradation in complex environments, hindering their commercialization. Recent developments in Large Language Models (LLMs) present new opportunities to address the above challenges. This study aims to explore the integration of LLMs into embodied AI systems, highlighting their potential to enhance scene understanding, reasoning, and planning capabilities. The paper provides a detailed review of LLMs' applications in embodied AI, demonstrating how these models can improve the robustness and adaptability of AI systems. Additionally, the study examines the limitations of LLMs, such as hallucinations and efficiency challenges, and discusses potential solutions to mitigate these issues. Through an in-depth analysis of LLM-powered enhancements in embodied AI, this research underscores the transformative impact of LLMs on intelligent systems. By addressing current limitations and implementing innovative solutions, LLMs can significantly advance the field of embodied AI, paving the way for more versatile and intelligent systems that can operate effectively in diverse real-world environments.

Keywords: Embodied AI; Large Language Models; Robustness and Adaptability.

1. Introduction

The convergence of artificial intelligence and robotics has led to the emergence of embodied Artificial Intelligence (AI), where intelligent systems interact with the physical world and operate around humans. These systems, equipped with sensors and actuators, are designed to perceive, reason about, and act in real-world environments, revolutionizing industries such as autonomous driving, healthcare, and household assistance. For example, autonomous vehicles leverage multi-modal sensory inputs, such as images and Light Detection and Ranging (LiDAR) scans, to autonomously drive alongside human-driven vehicles. Similarly, generalist robots use their vision systems to understand users' intents and assist users in daily life.

Embodied AI systems typically consist of modular architectures that include three key modules: perception, reasoning, and motion/trajectory planning. The perception module is trained to convert raw sensor inputs into symbolic representations of the world, such as surrounding objects' 3D-bounding boxes and semantic labels. The behavior reasoning module is trained to predict the future behavior of other dynamic agents, such as humans in the scene, to keep the embodied agent away from danger. Finally, the motion planning module aims to develop an efficient and safe motion plan to guide the embodied agent to complete the desired task. Unfortunately, each of these modules has its own bottlenecks: perception models struggle with missing important objects in the driving scene (e.g., a hole on the road), behavior reasoning models fail dramatically when the predicting agent behaves outside of the training data distribution, and the motion planning module is limited by the quality of its reward function. In the last few years, numerous studies have sought to address these limitations. For instance, LiDAR point clouds have been utilized as an additional signal to enhance perception models' performance in detecting long-range and small objects [1]. Additionally, Sun et

al. [2] explore learning reward functions from expert demonstrations through inverse reinforcement learning to improve the quality of motion planners. Yet, all the above attempts have poor generalization abilities in complex real-world environments.

Large Language Models (LLMs) [3] are high-capacity auto-regressive models trained with internet-scale data. They have demonstrated impressive in-context learning and common-sense reasoning abilities, increasingly addressing various bottlenecks across the autonomy stack of embodied agents. This paper explores the application of LLMs in embodied AI, focusing on their integration to enhance the performance, adaptability, and intelligence of these systems. This study examines how LLMs can improve embodied agents' perception models, reasoning and task planning models, and motion planning models, as well as their role as an integrated autonomy stack (Fig.1), byleveraging their advanced capabilities in common-sense reasoning. By incorporating LLMs, embodied AI systems can become more robust and versatile, effectively overcoming the limitations of traditional approaches.

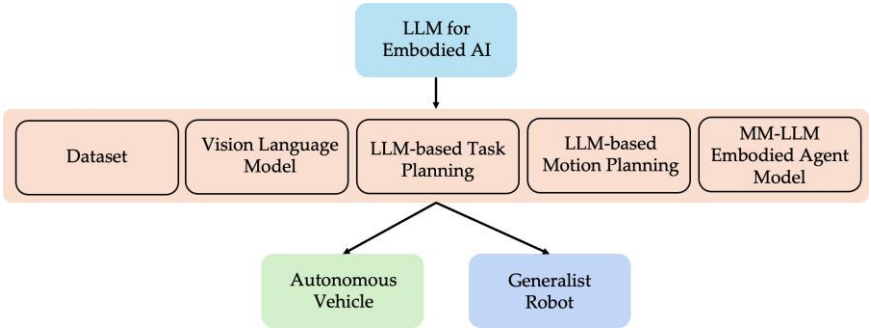

Figure 1. The Architecture of Study

This paper is structured as follows. Section 2 will introduce LLMs and current datasets available for developing LLMs-powered embodied agents. In Section 3, how LLMs can be used in embodied AI will be introduced. In Section 4,the limitations of LLMs for embodied AI and future opportunities to address these limitations will be discussed. Finally, Section 5 summarizes the paper and the conclusion.

2. Method

2.1. Dataset for Training LLM-Powered Embodied Agents

To effectively utilize LLMs in embodied tasks, fine-tuning with diverse Quality Assurance (QA) datasets specific to these tasks is crucial. For autonomous driving, the NuScenes1 dataset provides comprehensive sensor data and annotated scenes from Boston and Singapore, featuring multi-view images, LiDAR point clouds, and vehicle proprioceptive information. This enables the construction of QA pairs covering perception, reasoning, and planning, although most efforts focus on perception due to the ease of using ground truth information. Reasoning and planning QAs require costly human labeling, which limits their development [4]. In robot manipulation, available QA datasets are sparse and often collected in constrained environments, such as tabletop manipulation, making them insufficient for training large-scale models. Existing datasets typically contain fewer than one million QAs, posing a significant challenge for advancing robotics research. Corporate research institutes, like Google, create extensive datasets for their advanced LLMs, such as QAs from 96,000 training scenes for state-of-the-art robotics models [5]. However, these large-scale datasets are rarely open-sourced due to proprietary constraints, limiting their accessibility to the broader research community and hindering further advancements in the field.

2.2. Large Language Models (LLMs)

This section provides a brief introduction to LLM. LLMs are a class of deep generative models used to understand and generate text. LLMs, often comprising billions of trainable parameters, leverage

large-scale textual data to recognize the underlying patterns and structures of language. By capturing these intricate relationships, LLMs are able to generate human-like and contextually relevant text, even if it has not been seen during training. The core mechanism behind LLMs is learning the conditional text distribution, which enables the model to predict the likelihood of a word or phrase given its context. This allows the model to generate new, previously unseen text that is both grammatically correct and contextually appropriate. As a result, LLMs are able to perform a variety of natural language processing (NLP) related tasks with high proficiency, including chat-bot and summarization.

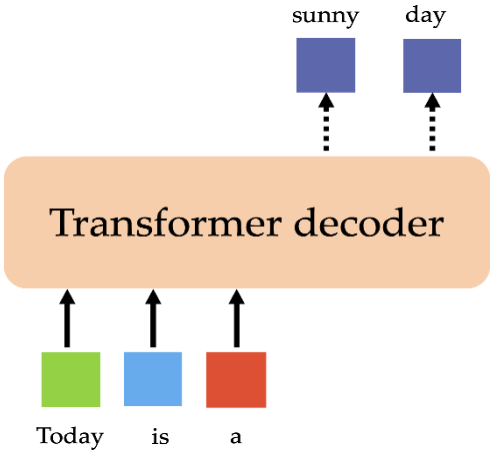

Figure 2. Illustration of The Transformer-decoder Model

LLMs typically use transformer-based models as their backbone, a class of neural networks that have revolutionized NLP and many other fields. These models are characterized by their use of attention mechanisms [3], which allows them to weigh the relevance of all parts of the input data simultaneously and more efficiently compared to the previously dominant recurrent neural networks (RNNs). A transformer-based model has an encoding model (encoder) and a decoding model (decoder). The encoder model is tasked with processing prompts into sets of feature representations, which the decoder model then uses to generate output sequentially. Recently, the decoder- only model becomes the predominant backbone for LLMs. The decoder-only transformer relies on stacked decoder layers, and each layer contains multi-head self-attention layers and feed-forward layers. One of the benefits of the self-attention layer is to enable the model to prioritize information from different parts of the input, dynamically adapting to the length and complexity of the input sequence (Fig.2). This model configuration allows it to focus on all preceding tokens to generate the next token in a sequence, effectively understanding and utilizing context without an encoder.

The most advanced LLMs include the Generative Pretrained Transformer (GPT) models developed by the company OpenAI [3]. While these models are highly advanced, OpenAI provides only access points for users to run inference and fine-tune the LLM, without full access to the underlying architecture. An alternative powerful open-source LLM is the Large Language Model Meta AI (denoted as LLaMA) model from Meta 2. Unlike GPT, LLaMA offers complete access to the model architecture and model card, allowing for extensive customization and application across various domains.

3. Applications of LLM in Embodied AI

An embodied agent is an AI system that utilizes multi-modal sensory inputs (e.g., images, LiDAR point clouds, haptic feedback, etc.) to interpret and act in the physical world. These agents fuse data from various sources to understand their environment, make decisions, and plan their actions. This section introduces traditional ap- proaches for designing embodied AI systems and explores the new opportunities unlocked by recent advancements in LLMs.

3.1. Traditional Approaches for Embodied Agents

Embodied AI systems typically consist three key modules: perception, reasoning, and motion planning. The perception module converts raw sensory inputs into symbolic representations, providing an abstracted and struc- tured understanding of the environment. The reasoning module then uses these abstracted representations to plan high-level decisions, determining the agent's goals and the necessary behavior plans. Finally, the motion planning module devises a safe and dynamically feasible trajectory for the agent to execute, ensuring precise and efficient movements for effective interaction with the physical world. Despite their success, traditional embodied AI systems face many challenges. These modular architectures often require extensive domain-specific knowledge and engineering to design and integrate the various components effectively. They can also be rigid and less adapt- able to dynamic and complex environments, limiting their flexibility and generalization capabilities. Additionally, inefficient interaction between modules can lead to suboptimal performance. These issues highlight the need for more integrated and adaptable approaches, where advancements in LLM can play a transformative role.

3.2. Vision Language Models as Perception Models

Vision-Language Models (VLMs) integrate visual and textual information, offering significant advantages over traditional perception models [6]. Unlike models that focus solely on visual data, VLMs combine insights from both visual and textual modalities to generate rich, contextual, and customized outputs. This integration allows VLMs to understand the nuanced context of images by leveraging accompanying text prompts, providing a deeper comprehension than models analyzing only visual data. The benefits of VLMs over traditional perception models include enhanced contextual understanding, enabling them to interpret complex scenes by combining visual and textual cues [7]. They possess open vocabulary detection abilities, identifying and understanding a wide array of objects based on the current context and concepts beyond those seen during training. By scaling up VLM training on large-scale diverse datasets, these models become more adaptable and flexible, capable of handling a variety of tasks and applications from autonomous driving to robotics with reduced dependency on extensive domain-specific training.

3.3. LLMs in Embodied Task Planning

Task planning is a crucial component of embodied AI systems, responsible for generating high-level behavior plans based on the agent's goals and environmental context. Traditional approaches to task planning have included rule-based systems [8] and reward function-based search [2]. In rule-based systems, the human designer encodes the task plan using a set of predefined rules. In reward function-based search, the human designer manually defines a reward function that describes the preferred way of executing the task and utilizes search-based algorithms to find an optimal behavior plan. Despite their success, these approaches are often limited by the designer's domain knowledge and inductive biases, which can restrict the system's adaptability and generalization to new scenarios.

LLMs are trained using internet-scale world knowledge, allowing them to implicitly encode semantic knowledge about the desired high-level behavior for completing tasks. This extensive training enables LLMs to understand and generate detailed, context-aware plans. As a result, many recent efforts have focused on utilizing LLMs as the task planning module in embodied AI systems. For example, LLMs can be used as a evaluation tool to score the probability that a low-level motion skill contributes to completing the instruction, or can be used to directly generate executable plans as programs [9].

3.4. LLMs for Motion Planning

Traditional learning-based planning models for embodied agents rely on symbolic and state-space perception outputs [2] to reduce the dimensionality of the inputs and accelerate the development cycle. However, these models typically require extensive human effort to design features, making them less robust to inductive bias and vulnerable to feature misalignment. An alternative paradigm is to train the embodied AI system end-to-end directly using multi-modality sensory inputs (e.g., images)[10].

Unfortunately, both paradigms suffer from severe performance degradation in long-tail scenarios in which training samples are extremely rare. To mitigate this issue, previous approaches mostly focused on detecting such situations online and switching to rule-based [2] planners to ensure safety.

LLMs are foundational models trained on world knowledge. In light of their outstanding understanding and generalization abilities, previous works have attempted to fine-tune LLMs into motion planners for embodied agents [11]. However, LLM-based planners are heavily dependent on the resolution and quality of the scene descriptions. While a more comprehensive and fine-grained description can help the LLM better understand the scene, it also makes the LLM inefficient to train and run inference. Additionally, designing templates to textualize scenes requires extensive prompt engineering.

3.5. Multi-Model Large Language Model as the Joint Perception & Planning Stack

Multi-Modal Large Language Models (MM-LLMs) tokenize multi-modal sensory inputs (e.g., image, force, audio, etc.) into latent tokens and inject these tokens into LLMs to help connect sensory inputs with text and have been increasingly used as an integrated autonomy stack that directly maps the sensory inputs to task-related outputs. The pre-dominant approach is to fine-tune existing MM-LLMs using various question-answer pairs from specific tasks. For example, [12] fine-tunes Flamingo using autonomous driving QAs to develop an MM-LLM model that can understand the driving scene and generate desired ego vehicle behaviors. However, due limited internet-scale embodied AI dataset. The models from this approach often lack the grounding, 3D geometry understanding, and behavior reasoning abilities essential for robotics. To mitigate this issue, recent works propose to pre-train the sensory inputs tokenizer using supervised learning on related sub-tasks (e.g., object-detection in autonomous driving, affordance prediction in robot manipulation [5]) to improve the 3-D grounding and planning capabilities of MM-LLMs.

4. Discussion

While LLMs offer significant advantages for embodied AI systems, they also present certain limitations. Two primary challenges are hallucination and efficiency.

LLM Hallucination refers to the generation of plausible but incorrect or nonsensical outputs. This occurs when the model produces text output that is not grounded in the provided context. In embodied AI systems, hallucination can lead to incorrect task plans and actions. For instance, an LLM might generate an action sequence that is not feasible in the physical world or suggest actions based on inaccurate interpretations of the environment. Hallucination can undermine the reliability and safety of the AI system, especially in critical applications like autonomous driving or robotic surgery, where precise and accurate actions are essential. To mitigate this issue, incorporating grounding mechanisms that tie the model's outputs to real-world data and context can help reduce hallucinations [13].

LLMs are very large and not efficient to train. Training these models involves processing massive datasets, which demands significant computational power, time, and energy. The size and complexity of LLMs result in extended training periods and substantial resource consumption. Additionally, LLM inference is slow due to its sequential nature. Generating responses or task plans requires processing tokens one at a time, which can introduce latency and reduce the responsiveness of the system. These inefficiencies have significant implications for embodied AI. The high computational demands can limit the deployment of LLMs in resource-constrained environments, such as mobile robots or embedded systems. The slow inference speed can lead to delays in task planning and execution, which is critical in dynamic environments where timely responses are crucial. These challenges can hinder the practical application of LLMs in real-time, mission-critical scenarios. To address the inefficiency in training, several techniques can be employed. Low-Rank Adaptation [14] is a method for reducing trainable parameters via efficient fine-tuning, thereby decreasing the computational load and speeding up the training process. Additionally, quantization techniques [15] can be used to accelerate training by reducing the precision of the model's calculations, which

decreases the computational requirements without significantly compromising performance. For improving inference speed, techniques such as caching can be utilized. By storing and reusing previously computed results, caching can reduce the amount of computation needed for each inference step, thereby speeding up the overall process. Additionally, model optimization techniques, such as pruning unnecessary parameters and using specialized hardware accelerators like Tensor Processing Units (TPUs), can further enhance the efficiency of LLM operations.

5. Conslusion

The convergence of artificial intelligence and robotics has led to the development of embodied AI, where intelligent systems interact with the physical world and operate alongside humans. Equipped with sensors and actuators, these systems are designed to perceive, reason about, and act in real-world environments, transforming industries such as autonomous driving, healthcare, and household assistance. Despite years of effort, embodied AI systems still suffer from limitations, including poor generalization ability and performance degradation in complex envi- ronments, hindering their commercialization. Recent advancements in LLMs offer new opportunities to address these challenges. This study examines the application of LLMs in embodied AI, highlighting their potential to enhance scene understanding, reasoning, and planning performance. LLMs provide transformative benefits for embodied AI systems, but addressing their limitations, such as hallucinations and efficiency challenges, is crucial to fully realizing their potential. Innovative solutions and careful implementation are needed to overcome these issues. By tackling these hurdles, LLMs can lead to more robust, versatile, and intelligent embodied AI systems, driving advancements across various domains. This study underscores the significance of integrating LLMs into embodied AI to achieve these improvements, ultimately contributing to the evolution of more capable and adaptive intelligent systems.

References

[1] Y. Zhou, P. Sun, Y. Zhang, D. Anguelov, J.Y. Gao, T. Ouyang, J. Guo, J.Q. Ngiam, and V. Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In Conference on Robot Learning, 1 (2020), 923–932.

[2] L.T. Sun, X.G. Jia, and A.D. Dragan. On complementing end-to-end human behavior predictors with planning. arXiv preprint:2103.05661 (2021).

[3] J. Achiam, S. Adler, S. Agarwal, et al. Gpt-4 technical report. arXiv preprint:2303.08774 (2023).

[4] X.P. Ding, J.H. Han, H. Xu, X.D. Liang, W. Zhang, and X.M. Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. arXiv preprint:2401.00988 (2024).

[5] D. Driess, F. Xia, M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint:2303.03378 (2023).

[6] C. Fei-Long, Z. Du-Zhen, H. Ming-Lun, C. Xiu-Yi, S. Jing, X. Shuang, and X. Bo. Vlp: A survey on vision-language pre-training. Machine Intelligence Research, 20(1) (2023), 38–56.

[7] K. Weicheng, C. Yin, G. Xiuye, A.J. Piergiovanni, and A. Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint:2209.15639 (2022).

[8] X. Wei, N. Mehdipour, A. Collin, A.Y. Bin-Nun, E. Frazzoli, R. Duintjer Tebbens, and C. Belta. Rule-based optimal control for autonomous driving. In Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems, 4(2021), 143–154.

[9] J. Liang, H. Wenlong, F. Xia, X. Peng, K. Hausman, B. Ichter, P. Florence, and A. Zeng. Code as policies: Language model programs for embodied control. In 2023 IEEE International Conference on Robotics and Automation, (2023), 9493–9500.

[10] H. Yihan, Y. Jiazhi, C. Li, L. Keyu, S. Chonghao, Z. Xizhou, C. Siqi, D. Senyao, L. Tianwei, W. Wenhai, et al. Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023), 17853–17862.

[11] M. Jiageng, Q. Yuxi, Z. Hang, and W. Yue. Gpt-driver: Learning to drive with gpt. arXiv preprint:2310.01415 (2023).

[12] M. Yingzi, C. Yulong, S. Jiachen, M. Pavone, and X. Chaowei. Dolphins: Multimodal language model for driving. arXiv preprint:2312.00438 (2023).

[13] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto. Multi-modal hallucination control by visual information grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2024), 14303–14312.

[14] J.H. Edward, S. Yelong, P. Wallis, Z. Allen-Zhu, L. Yuanzhi, W. Shean, L. Wang, and C. Weizhu. Lora: Low-rank adaptation of large language models. arXiv preprint:2106.09685 (2021).

[15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quan- tized llms. Advances in Neural Information Processing Systems, 36 (2024).