ISSN:2790-1688 Volume-15-(2025)

Large Language Models in Robotics: Applications, Challenges, and Future Directions from the Perspective of Embodied Intelligence

Ruijun Wang

School of Information Science and Technology (School of Cyber Security), Guangdong University of Foreign Studies, Guangzhou, 510006, China

Email: 20221003187@mail.gdufs.edu.cn

Abstract. Large language models(LLMs) are beginning to influence embodied intelligence by improving robots' ability to perceive their surroundings, follow instructions, and collaborate with people. This paper reviews recent progress at the intersection of LLMs and robotics, focusing on practical uses in manipulation, navigation, and human-robot interaction. At the same time, it highlights key difficulties, including grounding in the physical world, ensuring reliability and safety, and achieving efficient real-time operation. To address these gaps, the author discusses directions such as combining symbolic reasoning with neural methods and enabling robots to learn continuously through feedback. The study concludes that LLMs can strengthen embodied intelligence, but turning prototypes into robust real-world systems will require more efficient, adaptive, and human-aligned designs.

Keywords: Large Language Models, Embodied Intelligence, Robotics, Reliability, Human-Robot Interaction.

1. .Introduction

Recently, there have been breakthrough advancements in models, which are fundamentally reshaping the integration of artificial intelligence and robotics. Among them, some large language models play a key role, such as GPT-4, Claude, and Gemini. They have demonstrated relatively advanced capabilities in natural language understanding, reasoning, and a small amount of generalization[1][4]. Such capabilities have exceeded the limitations of traditional natural language processing applications. The synchronous development of embodied intelligence enables robots to perceive, reason and operate in real-world environments. Significant progress has been made in fields such as computer vision, operational strategies and reinforcement learning, which have greatly enhanced this ability of embodied intelligence. However, there is still a certain gap between high-level symbolic reasoning and low-level physical behavior at present. Correspondingly, there has been an increasingly obvious trend of integrating LLMs into robotic systems, as their advanced language understanding and reasoning capabilities can effectively bridge the gap between high-level decision-making and low-level physical execution of robots[5][6][7][8].

Embodied intelligence is the computational approach to the design and understanding of intelligent behavior in embodied and situated agents through the consideration of the strict coupling between the agent and its environment, mediated by the constraints of the agent's own body, perceptual and motor system, and brain[7][1][2]. Early approaches were dominated by classical control methods, such as PID controllers, but the increasing complexity of robotic tasks quickly demanded more adaptive solutions. Embodied intelligence, as a theoretical lens, goes beyond adaptive control by emphasizing intelligence from an agent's body-environment interaction; for LLM-powered robotics, it anchor LLM's abstract capabilities in physical actions to fit real constraints.

This review aims to synthesize current applications, identify challenges, and propose future directions for LLM-powered robotics through the lens of embodied intelligence. It first outlines how LLMs support robotics across four core areas – enhancing manipulation and tool use, optimizing navigation and embodied planning, facilitating human-robot collaboration, and boosting open-world generalization – then pinpoints key challenges including multimodal grounding and alignment, real-

time reasoning and computational efficiency, safety, reliability, and human alignment, and generalization and adaptation to open-world scenarios, before putting forward future research focuses such as developing data-efficient foundational methods, pursuing the core goals of "reducing computational overhead" and "ensuring real-time performance", deeply integrating formal verification methods and safety-constrained control layers into LLM-driven planning pipelines, and addressing the insufficient generalization of LLM-embodied systems.

2. Foundations of LLMs and Embodied Intelligence

LLMs, widely used Transformer-based framework today, are trained on massive corpora and exhibit strong generalization across reasoning and language tasks. GPT-4, PaLM, and LLaMA are typical cases here – their training covers text, code, even multimodal data, so they handle question answering, text summarization, and basic planning well [1][4]. Notably, they don't need task-specific fine-tuning to generate coherent outputs or do reasoning work. This trait makes them quite suitable for supporting autonomous agents, especially in areas that require parsing complex instructions or balancing multiple objectives.

Despite these linguistic and reasoning strengths, LLMs remain fundamentally disembodied- they work purely with symbolic representations of language and data, no direct links to sensorimotor experiences or physical environments. Meanwhile, they can describe how to grasp a cup or navigate a room, but they do not "feel" the weight of the cup or "see" the obstacles in the room, which might cause a mistake: a model may generate a possibly suitable instructions, but when these come to a robot, the execution would turn to a failure for the model has no grounding in sensory and motor realities. This limitations emphasizes the reason why many researchers think LLM is not sufficient enough for the robust embodied behavior alone[5][9][11].

To solve this limitations, embodied intelligence (EI) which is rooted in cognitive science and robotics provides a theoretical frame. In EI field, the intelligence comes from the interaction of cognition, action and perception[1][2]. Without a body interacts with the environments, reasoning made by LLMs cannot be fully understood or effectively deployed. From this point of view, EI means the robots must be combined with sensory inputs, physical actions and self-adaptive feedback loops to achieve robust and context-aware behaviors. This perspective also emphasizes the importance of grounding symbolic representations in perceptual and motor experiences which is the key to ensure computational abstract planning come to the real world[2][11].

The convergence of LLMs and robotics under the embodied intelligence paradigm shows a possible path forward. By combining the high-level reasoning capabilities of LLMs with the sensorimotor in EI, researchers are able to develop the agents that not only can understand and plan for the tasks, but also execute them reliably in the dynamic environments. For example, LLMs could generate multi steps of the tasks by natural language and transform to the robot to interpret it and execute what it needs to do by its cognition and motion to closing the gap between abstract reasoning and performance in physical environments. The integration establishes the possibility of more adaptive, autonomous and generalizable robotic system and laid the foundation of applications, challenges and future directions.

3. Applications of LLMs in Robotics

3.1 Manipulation and Tool Use

Manipulation has been a bottleneck for robotics research for a long time because it demands precise motor control and the ability to interpret task instructions flexibly. Recently, the use of LLMs in this area reflects a kind of transformation: researchers increasingly ask whether it is possible to help robots plan and adapt manipulation strategies by training general model on vast corpora instead of designing explicit controllers for every scenario[3][5].

One influential study was PaLM-SayCan[6], which showed that a language model could be paired with pre-trained low-level skills to decide "what to do next" in household work. The system was able to break down the vague request from users to a series of executable actions and choose the skills that were reliable from its skill pool[6][7]. RT-2 and other following works pushed this by training a visual-language-action model that could turn its knowledge into physical actions. This means that in the real world, robots can recognize objects not included in their training and attempt to manipulate them in a reasonable way[7][8].

Smaller-scale experiments reinforce the trend. Some groups have proved that LLMs can provide commonsense reasoning about tools, for example, suggesting using a spoon to scoop liquid or reaching a stick by hands. These demonstrations highlight the advantages of building the symbolic reasoning upon the big-scale knowledge base, which the pure reinforcement learning policies usually do not have[5][9].

However, this approach has clear shortcomings. LLM-based planners usually need external grounding modules to translate the language into motor commands. And they are still brittle in scenarios that involve dynamics, force control, or fine-grained physical interactions. Real-time performance is another challenge: it might take longer to use a large model for inference than what is acceptable in the embodied setting[12][13].

3.2 Navigation and Embodied Planning

Navigation and embodied planning belong to another key application area of LLM-driven methods. The core requirement of this field is to enable the robots to rely on environmental perception and task instructions to achieve autonomous path planning, target positioning and dynamic adjustment in the complex physical spaces. In this domain, traditional approaches relied on geometric mapping and path-planning algorithms. Though these algorithms work well in constrained environments, when using a high-level and ambiguous language for description they often fail. In contrast, LLMs can interpret this kind of instructions and reason the possible actions, which brings the robot more flexibility in unstructured or dynamic settings[8][14].

PaLM-E [8] was a typical representative in this area. It integrates a large language model with embodied sensory inputs. Different from the early cognitive and planning modules, PaLM-E accepted the visual observations and language queries which allowed the system to reason with text and cognition together. This trait brings the robot capability of obeying the instructions by generalized reasoning and spatial cognition, for example, "Find the fruit on the table in kitchen and take it to the dining table". To our surprise, this model could generalize to those navigation and manipulation tasks that have not been specifically programmed for it actually using web-scale knowledge to guide the unfamiliar layouts' behaviors[15].

Except for individual case studies, researchers have explored embedding LLMs in simulation frameworks like Habitat or iGibson, where the agents need to do numbers of navigation tasks with linguistic guidance. These environments make it possible to study how language-based reasoning supports subgoal generation. For example, they need to learn that "if you want to reach the office, you might go through the hall." It has proven that it is difficult to encode in traditional planners but it becomes easy when language models are used as high-level controls.

However, there are still bottlenecks. The performances relied on efficient grounding: if the visual inputs are noisy or incomplete, the planner's reasoning often breaks down. Another challenge is latency. Since the robots must react to the moving obstacles or dynamic human movements, models take lots of time in computing. What's more, the success of simulation does not always transfer to hardware for the additional difficulties from sensor noise, occlusion and clutter[8][11][13].

3.3 Human-Robot Collaboration

Among the application area of LLMs in robotics, human-robot collaboration is perhaps the most delicate. Different from single-agent planning, it requires robots to reason and act with human. In this situation, natural language is not only a communication medium, but also a coordination mechanism.

The key difficulty is how to make robots understand intentions expressed in language and synchronize them with own decision-making logic.

A representative example is the Inner Monologue framework [9]. Instead of hiding its internal working, the robot shows a continuous stream of clear language-based reasoning chains, which is consistent with the real-time feedback from sensors. There are two benefits: first, collaborators can directly see how the robot interprets the environment and why it takes such actions. Second, when the reasoning is visible, humans are more willing to let the robot take the responsibility of subgoals.

More importantly, the clear reasoning channel makes human intervention possible. If the robot misunderstands or makes mistakes in the instructions, the collaborators are able to step in and fix it by reading the reasoning logs. This kind of "understanding and intervening creates a new collaboration model. Robots are no longer just tools to execute instructions but have become communication partners and provide a more practical technical approach. From the perspective of embodied intelligence, this step is crucial because it ties together linguistic reasoning and physical interaction in real time.

3.4 Open-World Generalization

One of the hardest goals in connecting LLMs with robotics is to make systems that work outside the lab. Unlike benchmark tasks, the real world does not have clear boundaries. Objects may look quite different, instructions can be vague, and there are often some unexpected obstacles. In these cases, a robot has to make decisions without knowing what will happen.

A well-known attempt in this direction is SayCan[6]. The system links two elements: a language model that can interpret instructions and a grounding layer that can check what robot can actually do. When given an instruction like "bring me some snack", the LLM can break it into several steps, but the execution plan depends on whether the robot is physically able to keep going. In other words, the decision is made by a filter through a feasibility model. The design prevents the robot from making promises it cannot keep, which was a problem in the past when robots only relied on language.

The interesting part is SayCan would blend prior knowledge from language with the feedback from the environment. Which makes robot not just "follow the orders". It will keep updating its plan according to what it sees and what is possible. Even so, the approach is still limited: instructions in daily life can be underspecified or the environments can change too quickly. These gaps point to the real frontier: how to make robots generalize when both language and the environments around them are uncertain.

4. Challenges and Limitations

4.1 Grounding and Embodied Alignment

One of the central obstacles in combining LLMs with robotics is the gap between symbolic instructions and physical execution. Though a model can generate coherent action plans in language form, it is still fragile to ground them in sensorimotor reality. Perception noise, occlusion, and inaccurate object recognition often cause mismatches between linguistic tokens and actual entities. These errors accumulate in long-term tasks where misalignment at early steps can lead to a complete failure. Moreover, high-level priors from language sometimes override the immediate sensory inputs resulting in unsafe or inefficient behaviors. Systems like SayCan and PaLM-E tried to solve this problem by linking language subgoals with a library of verified skills, but the problem is far from completely solved. Reliable grounding will require tighter integration of perception, language and continuous feedback from the environments[11].

4.2 Safety, Reliability and Human Alignment

When robots act in human spaces, safety and reliability comes to the first place. Unlike software agents that only produce text, an error in physical world might break objects or even injure people. Current LLM-driven controllers inherit known issues such as hallucinations, vague reasoning and

opaque decision paths. These raise two major problems: users find it hard to believe behavior they cannot interpret and responsibility is unclear when failures occur – was it the model's faulty reasoning, the robot's low-level execution or ambiguous human input?

To address this problem, researchers are exploring layered safeguards: formal verification of planned actions, real-time constraint controllers as a final defense and explainable reasoning modules that expose why a robot chose a specific plan.Coupled with human-aligned protocols sensitive to moral norms and cultural context, these measures aim to reduce risks and support trustworthy deployment[9][11].

4.3 Real-Time Performance and Efficiency

Achieving real-time reasoning under strict computational limits is a major challenge for LLM-based robots. Most advanced models require massive parameters and cloud-scale resources, which do not match the limited processing power and energy budgets of mobile platforms. Thus, delays in inference reduce task efficiency and pose safety risks, particularly in dynamic environments where rapid adaptation is essential. Several technical directions aim to bridge this gap. Model compression, such as distillation and quantization, reduces parameter scale while preserving reasoning ability. Hybrid edge–cloud frameworks offload complex reasoning to the cloud while keeping time-critical responses local. Hardware accelerators and efficient architectures further shorten inference latency. The key challenge is balancing lightweight design with sufficient reasoning power to sustain scalability and robust task performance[16][17].

4.4 Generalization and Robustness

LLMs-driven robots often perform well in familiar or controlled environments but can struggle once the context changes. The system can be easily confused by new objects, vague instructions or complex layouts which can make the behavior unpredictable. Continuous learning provides a solution, where robots gradually pick up new knowledge while maintaining their existing skills[18]. Meta-learning offers another path: the robot can adjust to unfamiliar situations with minimal guidance by learning general patterns of adaptation. Human feedback can also help the system to refine its actions based on corrections. Such practical tools as sim-to-real adaptation and mechanisms that account for uncertainty further support consistent performance[19].

5. Future Directions from the perspective of Embodied Intelligence

5.1 Multimodal and Sensorimotor Integration

An important future direction is to involve integrating multiple sensory modalities into LLM-driven robots. Currently, many systems divide linguistic understanding and cognition, which limits their adaptation to the real-world environments. With the combination of visual, tactile and auditory inputs, robots can better understand the complex instructions. For example, robots can deal with a spoken instruction while analyzing the shapes, textures and context of the objects, which help them to make a more detailed plan for the coming actions. Achieving this integration required new architectures to merge different data streams and align them with symbolic reasoning[20][21].

5.2 Neuro-symbolic and Hybrid Approaches

Another promising approach is to merge the flexible reasoning of LLMs with the reliability of symbolic systems. Plans that are generated only by linguistic guidance might be unsafe or lack logic, especially in multi-steps tasks. By combining symbolic rules and formal logic into LLMs planning, robots would acquire the traits of adaptability and correctness. For example, a mix-system is able to use LLMs to suggest a series of actions and symbolic planners could check the viability and safety before executing the following actions. The further development of neuro-symbolic techniques could

make robots reason effectively in complex environments with high reliability and explainability[19][23].

5.3 Energy-Efficient and Real-Time LLMs for Robots

Running LLMs on robots is a big challenge due to the limitations of computing power and battery life on mobile platforms. Large models might slow down the performance and impact the efficiency and safety. One solution is to distillation or compression the model to reduce the size so that it can preserve the reasoning ability and be lightweight. Another possible way is to divide tasks between robots and cloud severs. The local agent could handle those simple and urgent decisions while the cloud one could deal with those complex things. The goal is to find a balance: models should be compact enough for quick execution while still capable of performing real-world tasks effectively[24][25].

5.4 Continual and Interactive Learning

Robots must keep learning over time for they need to operate effectively in real-world environments. Continual learning allows robots to adapt to new tasks, objects and environments without forgetting earlier skills or solely depending on pre-trained knowledge. Meta-learning can speed up adaptation by teaching robots general strategies and interactive learning introduce human guidance and feedback to help robot refine its behavior, correct mistakes and reshape preferences[20]. Over time, this interactive learning mechanism make robots better at understanding users' intentions and adapting to changing environments.

The directions outlined – multimodal integration, neuro-symbolic reasoning, efficiency optimization, and continual interactive learning – together offer a roadmap for building more capable and adaptable robots. Viewed through the lens of embodied intelligence, these approaches aim to bridge perception, reasoning and action, enabling systems that can respond flexibly to real-world challenges while maintaining reliability and safety.

6. Conclusion

In conclusion, this paper reviewed the progress and challenges of linking large language models with robots. The author found that while LLMs improve reasoning and task flexibility, they face limits in grounding, speed and reliability. From an embodied intelligence view, the key is not only language ability but also the link to perception and action. Without this bridge, robots stay strong in words but weak in practice. Looking forward, research should focus on lighter models, closer human-robot interaction and after deployment learning. These steps can bring robots from controlled tests into daily use with grater trust and safety.

References

[1] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159.

[2] Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

[3] OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[4] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in LMs. *Proceedings of the 58th ACL*, 5185–5198.

[5] Vemprala, A., Guo, D., Gopalakrishnan, K., & Kappler, D. (2023). ChatGPT for robotics: Design principles and model abilities. *arXiv preprint arXiv:2306.17582*.

[6] Ahn, M., Brohan, A., Chebotar, Y., Cortes, J., Francis, A., Fu, C., ... & Levine, S. (2022). Do as I can, not as I say: Grounding language in robotic affordances. *Robotics: Science and Systems (RSS)*.

[7] Brohan, A., Brown, N., Chebotar, Y., Finn, C., Gopalakrishnan, K., Hausman, K., ... & Levine, S. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

[8] Driess, D., Xia, F., Chowdhery, A., et al. (2023). PaLM-E: An embodied multimodal language model. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[9] Huang, J., Abbeel, P., Pathak, D., & Mordatch, I. (2022). Inner Monologue: Embodied reasoning through spoken language. *arXiv preprint arXiv:2207.05608*.

[10] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[11] Singh, A., Hill, F., Santoro, A., Botvinick, M., & Lake, B. (2023). Grounding large language models in interactive environments. *arXiv preprint arXiv:2302.02662*.

[12] Li, J., Xia, F., Chen, X., et al. (2022). Pre-trained language models for interactive decision-making. *NeurIPS*.

[13] Khandelwal, U., Wang, J., et al. (2023). Embodied language models. *arXiv preprint arXiv:2309.13007*.

[14] Gervet, T., Wang, C., et al. (2023). Navigating with LLMs: Natural language navigation and planning. *ICLR*.

[15] Zhu, Y., Xu, J., et al. (2023). Multimodal foundation models for embodied AI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[16] Chen, S., Liu, J., et al. (2023). Can large language models reason about robotics? *arXiv preprint arXiv:2304.11110*.

[17] Xu, K., Zhang, Y., et al. (2023). EmbodiedGPT: Vision-language pre-training via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2305.15021*.

[18] Duan, Y., Andrychowicz, M., Stadie, B., Ho, J., Schneider, J., Sutskever, I., Abbeel, P., & Zaremba, W. (2017). One-shot imitation learning. *NeurIPS*.

[19] Parisotto, E., & Salakhutdinov, R. (2021). Neuro-symbolic methods in reinforcement learning. *ICML Workshop*.

[20] Kirk, R., Yin, Y., et al. (2023). Continual learning with foundation models in robotics. *arXiv preprint arXiv:2308.12952*.

[21] Thrun, S., & Pratt, L. (1998). Learning to learn. *Springer*.

[22] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

[23] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. *MIT Press*.

[24] Otte, M., & Frazzoli, E. (2016). RRT-based path planning in dynamic environments: A survey. *Autonomous Robots, 42*(2), 1–19.

[25] Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research, 17*(39), 1–40.