Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 1, 2025

From Conversation to Action: Opportunities and Challenges of Large Language Models as the Brain of Humanoid Robots

Yiyang Shao

Department of Mechanical and Automation Engineering; The Chinese University of Hong Kong; Hong Kong 999077; China

Abstract: Humanoid robots, as the core carriers of embodied intelligence, are moving from specialized scenarios to general-purpose applications. The incorporation of large language models (LLMs) endows them with cognitive and decision-making capabilities similar to those of a human "brain," becoming a key enabler for technological breakthroughs. Based on the practical and technological achievements of the global humanoid robot industry from 2024 to 2025, this paper systematically analyzes the technical value and practical bottlenecks of LLMs as robot brains. The study finds that, through semantic parsing, task planning, and generalized learning capabilities, LLMs increase the success rate of task decomposition in complex scenarios to over 95%, reducing deployment costs by 60%. However, they also face core challenges such as data shortages, insufficient real-time performance, and poor architectural adaptability. Drawing on practical examples from Google's RT series, Tesla's Optimus, and domestic companies such as iFlytek and UBTECH, this paper proposes a three-step breakthrough path: "data accumulation - architecture optimization - ecosystem collaboration." Research indicates that the deep integration of LLM and robotics will accelerate the scale-up of the industry. The Chinese market is expected to reach 75 billion yuan in 2029, accounting for 32.7% of the global market.

Keywords: Large Language Model, Humanoid Robot, Embodied Intelligence, Generalization Ability, Data Bottleneck, Task Planning.

1. Introduction

At the 2024 World Artificial Intelligence Conference and the World Robotics Conference, humanoid robots equipped with large models were a hot topic. The Dahua XR4 robot, leveraging Robot GPT, was able to interact with complex environments, while iFlytek's humanoid robots were able to perform fine movements such as brewing coffee and wiping towels, achieving a two-fold improvement in motor performance compared to previous generations. This marks the humanoid robotics industry's transition from advancements in the "cerebellum" (motor control) to upgrades in the "brain" (cognitive decision-making). Traditional humanoid robots are limited by pre-programmed functions, capable of completing single tasks in fixed scenarios and unable to function outside of their pre-programmed environments. As Ji Chao, Chief Scientist of iFlytek Robotics, put it, previous robots "understood actions but not purpose," failing to integrate physical operations with business processes. Since 2023, the explosive growth of LLM technology has provided a solution to this dilemma. Its commonsense reasoning, multimodal understanding, and task decomposition capabilities enable robots to understand the semantics of their environment, parse complex commands, and autonomously plan action chains, truly achieving the leap from "dialogue response" to "autonomous action." The global humanoid robot market is currently expanding at an annual rate exceeding 20% and is expected to reach tens of billions of dollars by 2025. The Chinese market is expected to exceed 75 billion yuan by 2029, accounting for 32.7% of the global market. However, the integration of LLM and robotics is still in its infancy: leading domestic companies have only achieved Level 2 intelligence, with limited generalization capabilities and a sharp decline in the success rate of complex tasks. Against this backdrop, systematically examining the technical opportunities of LLM as the robot brain, analyzing the practical challenges in its implementation, and exploring feasible breakthrough paths are of great theoretical and practical significance for advancing humanoid robots from the laboratory to large-scale application.

2. The Technical Foundation and Core Logic of LLM Empowering the Robot Brain

The value of LLM, the brain of humanoid robots, lies in bridging the gap between human commands and physical actions through language understanding and logical reasoning. Its technical implementation relies on multimodal fusion and the construction of a closed-loop cognitive architecture. The integration of multimodal information is a prerequisite for LLM's effectiveness. Traditional robots rely on single-modal models: the visual system can only recognize object form, and the language module can only process text commands, failing to form scene cognition. LLM, however, integrates visual, tactile, and language data to achieve an integrated "perception-understanding-decision-making" process. For example, Google's RT-2 model, the first "vision-language-action" (VLA) model, can directly control a robotic arm using complex text commands, converting environmental images and verbal commands into specific action parameters. Bytedance's GR-2 model, pre-trained on 38 million internet video clips to learn human behavior patterns, has a 97.7% success rate in grasping special objects, such as transparent and soft objects. This multimodal fusion capability enables robots to "understand scenes and understand requests" just like humans. A closed-loop cognitive architecture forms the core of the robot's brain. This typical architecture consists of

five components: a perception layer, an object model layer, an LLM inference engine, a behavior planning layer, and an execution layer. Dynamic optimization is achieved through bidirectional data flow. When the robot receives the command to "water the succulents on the balcony," the LLM inference engine first parses the object associations of "succulents, flower pots, and watering amount." The object model layer then utilizes environmental maps and soil moisture data. The behavior planning layer generates an action chain: "avoid obstacles, retrieve the pot, and spray." After the execution layer completes the operation, it feeds the results back to the LLM for parameter correction. This closed-loop mechanism ensures action accuracy and environmental adaptability and is key to transforming the LLM from a "language model" to an "action brain." Co-optimization with reinforcement learning further enhances the brain's fine control capabilities. The EUREKA algorithm, proposed at ICLR 2024, uses the LLM to generate reward functions and outperforms human experts in 83% of tasks across 29 reinforcement learning environments, enabling a simulated shadow hand to perform fine manipulations such as rapid pen rotation [1]. This "LLM-generated reward function + reinforcement learning action training" model solves the difficulty of traditional robots in fine manipulation and lays the foundation for applications in complex scenarios.

3. Core Opportunities of LLM-Driven Humanoid Robot Development

The integration of LLM not only improves the intelligence of individual robots but also creates breakthrough opportunities in three dimensions: scenario adaptation, industrial efficiency, and business ecosystem, accelerating the large-scale deployment of humanoid robots. Improved scenario generalization capabilities have broken down application boundaries. Previous-generation robots required retraining data and programming for specific scenarios, but LLM, through few-shot learning and task decomposition capabilities, achieves cross-scenario adaptability. In open scenarios, iFlytek's humanoid robots can automatically decompose the command "clean up the study" into an action chain of "organize bookshelves → wipe the desk → put away stationery" without manual pre-programming. In the industrial sector, UBTECH robots can now simultaneously perform sorting, handling, and quality inspection tasks, achieving flexible "multi-functional" configuration. This generalization capability transforms robots from "specialized equipment" into "general-purpose tools," suitable for diverse scenarios such as home services, industrial manufacturing, and disaster relief. This leap in industrial efficiency has lowered the barrier to commercialization. LLM significantly reduces development costs and cycles by automatically generating control code and optimizing motion planning. Data shows that the deployment cost of new LLM-based robot scenarios has decreased by over 60%, and development labor costs by 90%. 70% of the control code can be automatically generated by LLM. This efficiency improvement is even more significant in collaborative scenarios: nearly 20 UBTECH robots collaborate through a "group brain network" to complete the entire process from material warehousing to sorting, achieving an overall efficiency improvement of 40% compared to single-unit operations [2]. Furthermore, LLM-driven predictive services can proactively respond to needs, such as proactively delivering first aid kits by analyzing behavioral signals of the elderly, further expanding the application value of robots. The restructuring of the business ecosystem has opened up a trillion-dollar market. The scalability of LLM has spawned innovative models such as "Robot as a Service" (RaaS). Nursing homes can subscribe to a "dementia care package" that includes an anti-lost object model and an LLM module for emotion recognition. At the same time, the emergence of object model markets and skill stores has enabled hospitals, logistics, and other industries to share professional models and training data, creating a virtuous cycle of "data-model-application." At the policy level, the Ministry of Industry and Information Technology's "Guiding Opinions on the Innovation and Development of Humanoid Robots" explicitly mandates key technological breakthroughs in the "brain, cerebellum, and limbs" by 2025, providing policy support for industrial development. Driven by both market and policy, humanoid robots are expected to become the next generation of intelligent devices, following smartphones.

4. Practical Challenges and Technical Bottlenecks of LLM as the Robot Brain

Despite significant opportunities, LLM integration into humanoid robots still faces three core challenges: data, performance, and architecture. These bottlenecks directly hinder its progress from the laboratory to practical application. Data shortage and insufficient quality are the most prominent industry pain points. Humanoid robots adhere to the "scaling law," meaning that improving model capabilities is highly dependent on data volume. However, currently available data falls far short of generalization requirements. Tsinghua University researcher Su Hang noted that the humanoid robot data available online falls two to three orders of magnitude short of generalization requirements, and even with the training data volume of GPT-3.5, there is still a gap. The difficulty in data collection stems from the complexity of scenarios. In the real physical world, the diversity of object shapes, ambient lighting, and task processes leads to high data labeling costs and long processing times. Synthetic data is often out of touch with reality, making it ineffective for model training. UBTECH's Chief Brand Officer, Tan Min, admitted that replacing core positions would require five to ten years of real-world data accumulation and tens of billions of yuan in investment. Inadequate real-time performance and reliability make it difficult to meet practical needs. Interactions in the physical world require robots to respond quickly, but LLM suffers from significant inference latency issues: Google's RT-2 model has an inference frequency of only 1-5Hz and an output motion frequency of 1-3Hz, resulting in a latency of 0.3-1 second. Figure robots, on the other hand, experience latency as high as 2-3 seconds. This delayed response is particularly dangerous in emergency situations, such as disaster relief, where optimal operational opportunities may be missed [3]. Reliability is also a serious issue. Currently, most robots are at Level 2 intelligence, capable of completing pre-set commands in specific scenarios. When faced with unseen objects or unexpected situations, they are prone to misjudging attributes and repeating ineffective actions, significantly reducing mission success rates. Model architecture and hardware compatibility suffer from inherent flaws. Existing LLMs are mostly designed for language understanding, lacking a native understanding of the

laws of the physical world and struggling to process physical parameters such as force control thresholds and motion trajectories. Yushu Technology CEO Wang Xingxing noted that the industry's excessive focus on data has neglected the unification and optimization of large-scale embodied intelligence model architectures, resulting in difficulties in effectively utilizing data even when it is available. On the hardware side, LLM's high computing power requirements conflict with the computing power limitations of robotic terminals. Although Tesla's Optimus uses the same computing chips used in cars, it still compromises performance when processing multimodal real-time data. This "software-hardware" mismatch is a key obstacle to unleashing the power of the robot brain.

5. Domestic and International Case Studies of LLM and Humanoid Robot Integration

The practical explorations of domestic and international technology companies and research institutions have provided valuable experience in empowering robot brains with LLMs, demonstrating both the potential of the technology and the challenges of implementation, providing practical references for industry development. Internationally, the technical approaches of Google and Tesla are the most representative. Google has built a comprehensive technology ecosystem through a combination of model iteration and data accumulation. From SayCan, which pioneered the integration of language and physical feasibility, to PaLM-E, which integrates vision and language, and finally RT-2, which can directly output action commands, Google has gradually upgraded its task decomposition capabilities. Its latest RT-X series integrates a dataset covering 22 robot types and 527 skills, increasing task success rates by three times that of previous models. Tesla is pursuing an "end-to-end algorithm" approach, migrating the technology from its FSD V12 autonomous driving system to Optimus. This approach replaces traditional code with pure neural network control, reducing manual intervention and enabling technology reuse between the vehicle and the robot [4]. OpenAI's collaboration with Figure AI focuses on practical scenarios. The Figure 01 robot, leveraging LLM, has achieved material handling and equipment inspection in industrial environments, demonstrating initial commercial potential. Domestic companies have established a development model that combines platform empowerment with breakthrough scenarios. iFlytek, with the Spark large-scale model at its core, has built a robotics super-brain platform. It has empowered 420 robotics companies, connected 15,000 developers, and collaborated with UBTECH and Yushu Technology to explore multi-person, multimodal interaction solutions. Its 1.7-meter humanoid robot, on display, achieved a success rate exceeding 95% in disassembling complex tasks, capable of naturally performing actions like pouring coffee and wiping sweat. Tiangong 2.0 and Tianyi 2.0, both from the Beijing Humanoid Robot Innovation Center, leverage embodied multimodal large-scale models to enhance scene understanding, enabling industrial parts sorting and power inspection, respectively, significantly improving their "brain" capabilities compared to earlier versions. Dahua's XR4 robot, exhibited at the Shanghai AI Conference, is equipped with Robot GPT, enabling autonomous navigation and interaction in unstructured environments. UBTECH, through its "swarm brain network" technology, has enabled the collaborative operation of nearly 20 robots, demonstrating the value of LLM in swarm intelligence. These practices demonstrate the feasibility of LLM empowerment, but also highlight common challenges. Whether it's Google RT-2's low 30% success rate in complex scenarios or the urgent need for real-world data in domestic robotics, both demonstrate that technological breakthroughs still require focusing on two core areas: data accumulation and architecture optimization [5].

6. Exploring Pathways to Overcome Bottlenecks and Future Development Outlook

To address the current challenges of LLM-enabled humanoid robot brains, collaborative solutions must be built from three perspectives: data, technology, and ecosystems, driving the technology from "usability" to "efficiency." This data dilemma can be overcome through "virtual-reality integration + crowdsourcing": On the virtual side, digital twin technology can be used to build highly realistic virtual environments, recreating scenes like homes and factories to generate simulated operational data, significantly reducing the cost of real-world data collection. On the physical side, a crowdsourcing model using AR glasses is employed, encouraging users to annotate object attributes and scenario rules in natural language during daily interactions, such as "Grandpa's hearing aid needs to be charged at 8:00 every day," rapidly expanding the amount of data. At the same time, drawing on UBTECH's "Crowd Brain Network" technology, a cross-enterprise data sharing platform will be established to synchronize robotic operation experience across the enterprise, achieving "one-time learning, global reuse." The Ministry of Industry and Information Technology will lead the development of unified data labeling standards to address formatting issues and improve data utilization efficiency.

Technical architecture optimization should focus on "software-hardware collaboration + algorithm innovation": On the software side, a dedicated embodied intelligence model will be developed to enhance understanding of the laws of the physical world. At the same time, drawing on the EUREKA algorithm, efficient reward functions will be designed using LLM to improve reinforcement learning efficiency and address the challenges of fine manipulation. On the hardware side, high-computing, low-power devices will be developed, equipped with specialized chips such as the Tesla DOJO D1 and Texas Instruments TDA4x. This will create a collaborative "cloud-edge" system. The cloud focuses on model training and updating, while the edge handles real-time action decisions, reducing response latency to less than 0.1 seconds. A closed-loop "perception-reasoning-execution" mechanism will also be established. For example, if a robot cracks an egg after grasping it, it can automatically adjust force control parameters based on failure feedback to improve operational accuracy. The construction of the industrial ecosystem needs to rely on policy guidance and collaboration among multiple entities: the government can strengthen its support for open source platforms, launch open source projects such as the "Basic Model of Home Service Robots", and attract developers to contribute data and algorithms; implement the "Guiding Opinions on the Innovation and Development of Humanoid Robots" and establish special funds to support data collection and model development [6]. Enterprises need to explore diversified business models, such

as iFlytek's platform empowerment and UBTECH's scenario solutions, to dilute costs through large-scale applications. Universities and research institutions need to strengthen interdisciplinary cooperation in robotics, cognitive science, etc. A joint effort to build a "Human Life Object Graph" encompassing over 100,000 entities will clarify implicit rules within scenarios and supplement common-sense knowledge for the LLM.

Over the next 5-10 years, as technology matures, humanoid robots will gradually transition from industrial assistants to life partners, playing a role in flexible factory production, home care, disaster relief, and other scenarios. When the object model library exceeds 100 million entities, the LLM is expected to possess cross-domain innovation capabilities and independently develop emergency solutions, truly achieving the intelligent leap from "conversational understanding" to "action implementation."

7. Conclusion

The Large Language Model equips humanoid robots with an "intelligent brain," signaling the accelerated arrival of the era of embodied intelligence. This article, by reviewing industry practices and technological achievements from 2024 to 2025, clearly demonstrates the core value of LLM in semantic parsing, task planning, and generalized learning. It enables robots to transcend the limitations of traditional pre-programmed systems, increasing the success rate of task decomposition in complex scenarios to over 95%, reducing deployment costs by 60%, and driving the global humanoid robot market to expand at an annual rate exceeding 20%. China is expected to capture 32.7% of the global market share by 2029. However, the development of LLM as the robot brain still faces three bottlenecks: data scale is 2-3 orders of magnitude short of generalization requirements, inference latency of 0.3-3 seconds is insufficient to meet real-time requirements, and the model architecture is insufficiently compatible with the physical world. These issues have resulted in current robot intelligence levels remaining at only Level 2, and the success rate of complex tasks has plummeted. The practices of Google's RT series, Tesla's Optimus, and domestic companies such as iFlytek and UBTECH demonstrate that overcoming these bottlenecks requires the coordinated development of "data, technology, and ecosystem." The core solution lies in overcoming data shortages through virtual-real integration and crowdsourcing collaboration, improving real-time reliability through software-hardware collaboration and algorithmic innovation, and building an industrial ecosystem through policy guidance and multi-stakeholder participation. With the implementation of these measures, it is expected that embodied intelligence large-scale models will achieve architectural breakthroughs in the next two to three years, and reach a peak similar to ChatGPT within five years. Essentially, the integration of LLM and humanoid robots is not only a technological evolution but also a revolution in human-computer interaction—transforming robots from "command executors" to "intelligent collaborators." This transformation will not only spawn trillion-dollar new industries but also reshape production and lifestyles, driving the development of new-quality productivity. While many challenges remain, the direction of technological evolution is clear. With continued advancements in data accumulation and architectural optimization, humanoid robots will ultimately achieve the true leap from "speaking eloquently" to "integrating knowledge and action."

References

[1] Liu Ziyi. Research on intelligent agent decision-making based on deep reinforcement learning and large language model [D]. Nankai University, 2024. DOI: 10.27254/d.cnki.gnkau.2024.000075.

[2] Wang Jian, Shi E, Hu Hua, et al. Large language model for robots: opportunities, challenges and prospects [J]. Journal of Automation and Intelligence, 2025, 4(1): 52-64.

[3] Zeng Kai, Wang Yaonan, Tan Haoran, et al. Technology and prospects of embodied intelligent humanoid robots driven by AI large models [J]. Chinese Science: Information Sciences, 2025, 55(05): 967-992. DOI: CNKI: SUN: PZKX.0.2025-05-001.

[4] Zeng Feng, Gan Wei, Wang Yong, et al. A review of large-scale language models for robots [J]. arxiv preprint arxiv: 2311.07226, 2023.

[5] Cao Lin. Artificial intelligence robots and humanoid artificial intelligence: review, perspectives and directions [J]. arxiv preprint arxiv: 2405.15775, 2024.

[6] Bai Chenjia, Xu Huazhe, Li Xuelong. Embodied intelligence driven by large models: development and challenges [J]. Science China: Information Science, 2024, 54(09):2035-2082.