Received: 21 April 2025 | Revised: 16 June 2025 | Accepted: 17 July 2025 | Published online: 6 August 2025

REVIEW

Smart Wearable Technology
2025, Vol. 00(00) 1–16
DOI: 10.47852/bonviewSWT52025965

BON VIEW PUBLISHING

Frontier Exploration of the Fusion of Embodied Intelligence and Large Language Models

Fujiang Yuan[1], Hong Jiang[1], Haoran Guan[2] and Yanhong Peng[1,3,*]

[1]*College of Mechanical Engineering, Chongqing University of Technology, China*

[2]*Department of Mechanical Engineering, National University of Singapore, Singapore*

[3]*Department of Information and Communication Engineering, Graduate School of Engineering, Nagoya University, Japan*

Abstract: Embodied intelligence (EI) generates intelligent behaviors through interactions between the body and the environment but exhibits deficiencies in areas such as semantic understanding. In contrast, large language models (LLMs) possess powerful language processing capabilities. Integrating the two can enable a unified system that combines perception, cognition, and decision-making. This article first introduces the concepts, challenges, and recent breakthroughs in EI. It then analyzes the origin, core technologies, and research progress of LLMs. Subsequently, it explores the integration pathways and application scenarios of combining the two paradigms. Finally, the paper highlights ongoing challenges, such as the need for high-quality data and standardized benchmark environments, and underscores the importance of cross-industry collaboration as well as partnerships between academia and industry. The integration of EI and large language models represents a promising direction for the development of general artificial intelligence and is expected to drive innovative applications of intelligent systems across diverse domains.

Keywords: embodied intelligence, large language models, AI, frontier exploration

1. Introduction

Embodied intelligence (EI) is an intelligent system that perceives and acts based on the physical body [1]. It emphasizes that the intelligent agent does not exist in isolation but acquires information, understands problems, makes decisions, and performs corresponding actions through continuous and dynamic interaction with the surrounding environment, ultimately generating intelligent behavior and showing adaptability to the environment [2]. Simply put, the core of EI is that intelligence not only exists in the abstract thinking of the brain, but is also reflected through the actual interaction between the body and the environment [3].

EI has four core elements: ontology, intelligent agent, data, and learning and evolution architecture. Among them, ontology is the physical entity that carries intelligence, usually manifested as a robot with various forms, such as quadruped robots, composite robots, or humanoid robots [4]. Ontology has the ability to perceive the environment and move and perform operations and is the key medium connecting the digital world and the physical world. Its ability boundary directly determines the scope of executable tasks of the intelligent agent, so the design of the ontology with broad adaptability and generalization is crucial. Intelligent agents are the core of EI, deployed on the ontology, responsible for perception, semantic understanding, task decision-making, and control. With the continuous evolution of deep learning, intelligent agents increasingly rely on deep network architectures based on large language models (LLMs) and integrate multimodal perception capabilities (such as vision, speech, touch, etc.) to form a new generation of models with general perception and decision-making capabilities [5]. Intelligent agents can not only understand complex semantic tasks but also adjust strategies in real time in dynamic environments to complete high-level task goals [6]. At the same time, in order to cope with different task requirements, intelligent agents are designed in various forms, covering decision-making and control mechanisms of different modalities and complexity levels. Data is the key to driving the generalization ability of EI [7]. Unlike traditional LLM, which mainly relies on large-scale Internet data, the data required for EI is often more scarce and expensive and needs to be highly consistent with the perception and behavior interaction in the real environment [8]. Especially in the face of changing task chains and decision-making processes, data must not only cover a wide range of scenarios and environments but also have high-quality and high-semantic information expression capabilities. Customized and highly reliable data resources in industry application scenarios will become the core foundation for the successful deployment and implementation of intelligent agents in the future. Finally, the learning and evolution architecture is the key guarantee for EI to achieve adaptive and self-evolution

*Corresponding author: Yanhong Peng, College of Mechanical Engineering, Chongqing University of Technology, China, and Department of Information and Communication Engineering, Graduate School of Engineering, Nagoya University, Japan. Email: yhpeng@nagoya-u.jp

capabilities. Intelligent agents continuously optimize their own models through repeated interactions with virtual or real environments, learn new skills, strengthen strategies, and evolve more efficient problem-solving methods. In practice, virtual simulation environments, such as NVIDIA's Omniverse platform, provide an efficient and low-cost learning scenario, which effectively accelerates the evolution of intelligent agents in virtual space. However, the complexity of the real world is much higher than that of the simulation environment. How to achieve efficient migration and integration from simulation to reality (Sim2Real) has become an important research direction in current architecture design [9].

Although EI performs well in perception and action execution, it still has significant deficiencies in semantic understanding, abstract reasoning, task planning, generalization, and human–computer interaction. In particular, it lacks the ability to express complex tasks in language and high-level semantic control interfaces. At the same time, the data required to train embodied systems is expensive and difficult to obtain, which limits their large-scale deployment and generalization capabilities. With its powerful language understanding, knowledge transfer, multistep reasoning, and task decomposition capabilities, the LLM [10] can provide semantic support, strategy guidance, and natural interaction interfaces for EI, effectively making up for its shortcomings at the cognitive and language levels and promoting the evolution of EI toward a more general and efficient intelligent system [11].

In recent years, the LLM has become a hot topic in the field of artificial intelligence (AI) [12]. It is a complex system based on deep learning and natural language processing (NLP) technology, with billions or even hundreds of billions of parameters, which enables it to understand and generate human language [13]. The ability of this model is not limited to processing text but also has made breakthroughs in the direction of large multimodal models, successfully combining multiple modes such as text, images, and audio. Since 2018, LLM has demonstrated its outstanding capabilities in many fields, not only making progress in text generation but also making remarkable achievements in image generation and multimodal applications.

At the core of LLM is the Transformer architecture, which was proposed by Google in 2017. It can deeply understand the grammar, word meaning, and context of a language by learning from a large amount of text data. During the training phase, LLM absorbs a large text data set to grasp the laws and patterns of the language. As the model parameters increase, these models can more accurately predict the next word or generate a coherent text sequence. Today, the application scope of LLM has far exceeded the traditional fields of text generation, question-answering systems, machine translation, and text summarization [14]. They are now also showing potential in fields such as media convergence, digital transformation, education, and multilingual community communication, bringing innovation and transformation possibilities to these fields. Combining LLMs with EI can achieve the integration of perception, cognition, and decision-making, and give the intelligent agent a closed-loop capability from "understanding language" to "performing actions" [15]. This fusion not only improves the system's versatility and task transfer capabilities but also significantly reduces the demand for training samples. The embodied system can be controlled at a high level through natural language, making human–computer interaction more natural and efficient [16]. At the same time, LLM can be used as a strategy generator in reinforcement learning to help the embodied body explore and learn efficiently. In addition, LLM has excellent multimodal fusion capabilities and can integrate multiple inputs such as language, vision, and action, providing intelligent agents with contextual perception and task decomposition capabilities [17]. Overall, the combination of LLM and EI greatly enhances the flexibility, adaptability, and generalization of intelligent agents and is an important path to stronger general artificial intelligence.

Nowadays, the combination of EI and the grand prophecy model is widely used in various fields, but there are no articles summarizing and evaluating this aspect [18]. This article will summarize the relevant research on the combination of EI and LLM and provide application references for subsequent related researchers. The first section of this article mainly introduces the introduction of EI related technologies; the second section mainly introduces the origin of the grand prophecy model and related technologies; the third section introduces the relevant application fields of the combination of EI and LLM; the fourth section looks forward to the future of EI and LLM; and the fifth section summarizes this article.

2. Foundations of Embodied Intelligence

In this section, the relevant concepts of EI will be introduced, along with the difficulties associated with EI and the current research and development efforts in the field of EI.

2.1. Introduction to concepts related to embodied intelligence

In this section, we will introduce several related concepts around EI, which are embodiment, EI, disembodied AI, embodied intelligent robots, and embedded tasks.

Figure 1 shows the complete technical chain of the embodied intelligent system from sensory input to behavioral decision-making, as well as the synergistic relationship between related disciplines and technologies. The diagram divides the entire system into multiple levels: the left side is based on machine learning, social learning, and psychology as the theoretical support for intelligent behavior; the middle part includes hardware perception modules such as robot chip design, computer vision, speech recognition, force and tactile sensors, which constitute the perception front end for the interaction between the intelligent body and the external environment; and the right side shows the software system from sensory information to cognitive modeling and then to high-level task planning and control, which integrates interdisciplinary technologies such as NLP, graphics, physical simulation, and bio-medicine.

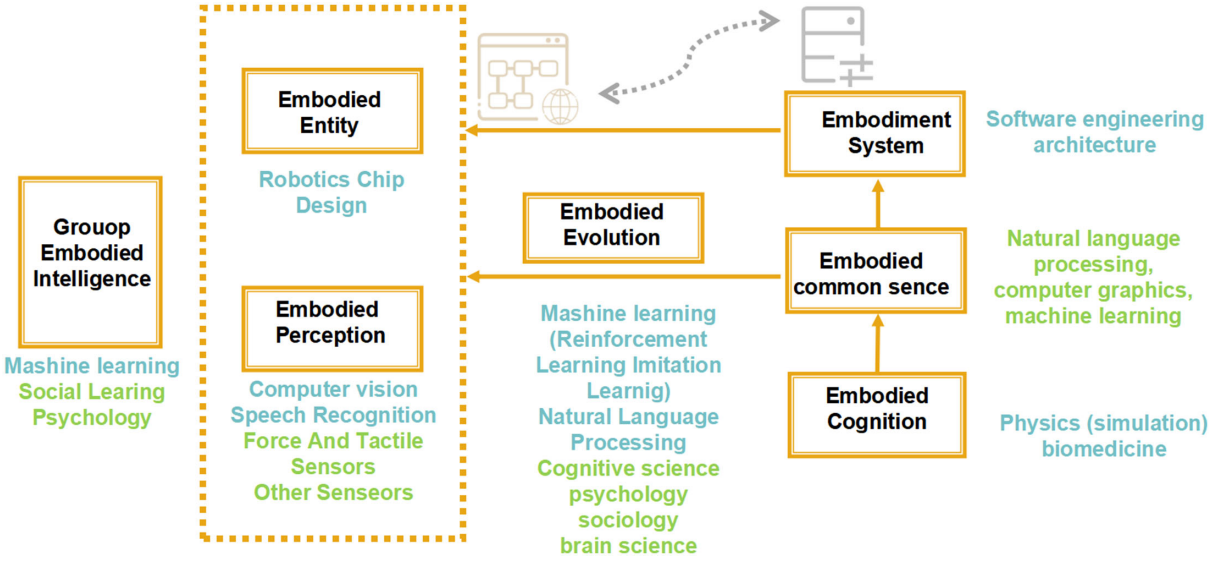
2.1.1. Embodiment and embodied

Embodied means that an intelligent agent has a physical body that can support sensing and motor control. This body is not only a morphological structure but also the basis for the intelligent agent to interact with the real world [19]. It enables the intelligent agent to affect the environment through its own actions and perceive the changes brought about by these actions, thereby achieving closed-loop learning and adaptation. Embodied means that a system or intelligent agent has a physical form and can perceive and interact. An "embodied" system not only exists in physical space but also can actively participate in environmental interactions, such as identifying objects, operating tools, and navigating to avoid obstacles [20]. The existence of this "physicality" is the key to the intelligent agent's true autonomous behavior.

2.1.2. Embodied AI and disembodied AI

Embodied AI refers to AI systems that not only have cognitive abilities but also have bodies and can actively interact with the

Figure 1
Components of embodied intelligence



environment. Unlike traditional models trained with static data, EI emphasizes learning through interaction with the environment. Its representative forms include service robots, self-driving cars, bionic robots, etc. [21]. These systems can "act physically" like humans, continuously optimize strategies through observation, action, and feedback, and thus complete complex tasks. Specifically, embodied intelligent agents need to have the following capabilities:

1) Language understanding: understanding human language instructions or dialogues.
2) Task planning and decomposition: being able to automatically decompose specific subtasks from high-level language commands.
3) Perception and positioning: identifying and locating target objects in the physical environment.
4) Action and navigation: performing actions, such as moving, grabbing, and manipulating objects.
5) Feedback and adaptation: adjusting action strategies in real time based on environmental feedback.

Disembodied AI is a representative of current mainstream AI models, such as LLMs and image recognition models, which usually rely on data collected and labeled by humans in advance for training [22]. This type of AI does not have the ability to physically interact with the real environment, so its intelligence is more inclined to "paper talk" or "strategic" reasoning. Although it can complete tasks such as language understanding and image generation in virtual space, it has poor generalization ability when facing real-world problems and lacks the ability to act autonomously and adapt to changes in the real environment.

2.1.3. Embodied intelligent robots

Embodied intelligent robots are typical representatives of embodied AI, with the ability to perceive, understand, act, and learn. They usually have the following characteristics:

1) Multimodal perception: perceive multimodal information such as images, sounds, languages, temperature, vibrations, etc., through cameras, microphones, force tactile sensors, etc. [23].
2) Semantic understanding and reasoning: understand complex instructions and make decisions based on environmental information.
3) Movement ability: able to navigate in space, bypass obstacles, grab and manipulate objects, etc.
4) Interactive learning: acquire new knowledge through interaction with the environment, achieve self-optimization, and enhance learning [24].

These robots can not only perform repetitive labor (such as sweeping robots) but also undertake more complex household services, medical assistance, warehouse management, post-disaster search and rescue, etc., becoming truly "useful and smart" physical intelligent bodies. Figure 2 shows the perception and interaction system of an embodied intelligent robot.

Figure 2 shows the robot's perception and interaction system. Computer vision gives the robot visual capabilities, just like humans use their eyes to observe the world; acoustic sensors enable the robot to hear, so that it can receive sound information; chemical sensors simulate the sense of smell and can detect environmental chemicals; natural language understanding and interaction modules allow the robot to understand and respond to human language; and tactile sensors allow the robot to perceive physical stimuli such as pressure, temperature, and fluid. These systems work together to help the robot perceive and interact with the outside world.

2.1.4. Embodied tasks

Embodied tasks refer to those tasks that require the intelligent agent to observe, move, speak, and interact through physical participation like humans. These tasks not only test the perception and understanding ability of the intelligent agent but also require it to have flexible response capabilities in complex environments [25]. Interactive teaching or collaborative tasks: such as assisting the elderly to take medicine, assembling equipment in collaboration with humans, etc. [26]. This type of task usually has the characteristics of strong environmental uncertainty, vague goals, and complex feedback and is an important criterion for evaluating the actual capabilities of embodied intelligent systems [27].

As shown in Table 1 above, Internet AI and embodied AI have essential differences in data sources, learning methods, perception capabilities, action capabilities, generalization capabilities, and

Figure 2
Perception and interaction system of embodied intelligent robots

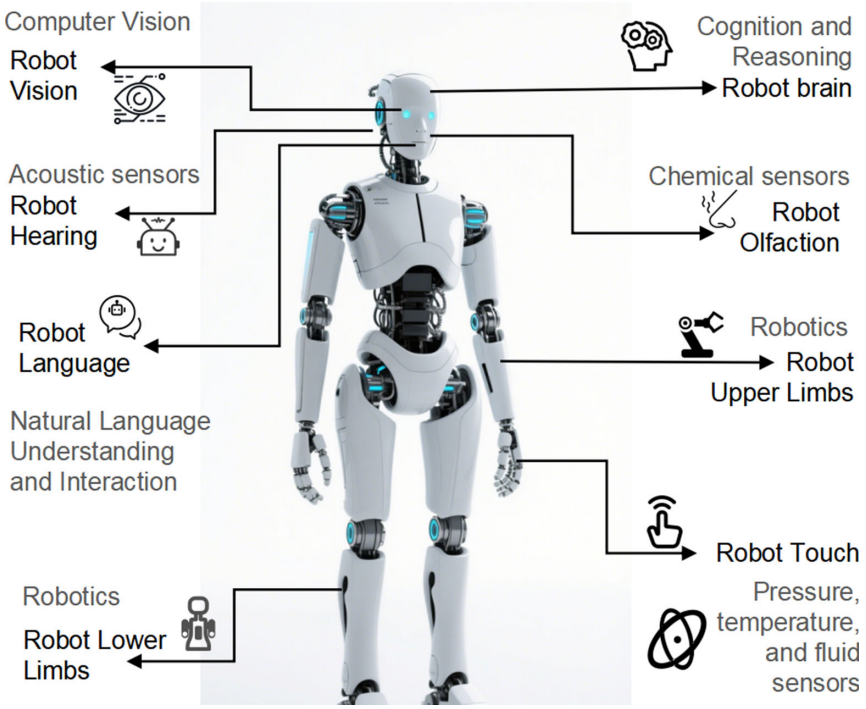


Table 1
Comparison: Internet AI vs. embodied AI

Feature	Internet AI	Embodied AI
Data Source	Static datasets (e.g., images, text, videos)	Real-time sensory data from dynamic interaction with the environment
Learning Method	Learning from human-collected/annotated data	Learning through interaction with the environment; self-learning
Perception	Usually unimodal or weakly integrated	Multimodal integration with comprehensive perception
Action Ability	No physical actuation/control, unable to affect the environment	Capable of sensing, moving, manipulating, completing task loops
Generalization	Weak generalization, prone to overfitting specific tasks	Stronger adaptability to real environments, better generalization
Intelligence Form	Strong in modeling, reasoning, and generation, but lacks real-world experience	Embodied intelligence combining understanding with action

intelligent manifestation. Internet AI mainly relies on static data sets for training, usually in a single-modal or weakly fused form. Although it performs well in modeling and reasoning generation, it lacks interaction with the real world, resulting in poor generalization and a lack of action capabilities. Embodied AI, on the other hand, obtains multimodal perception data through real-time interaction with the environment and has physical behavioral capabilities such as perception, movement, and operation. It can self-learn and adapt in the actual environment, reflecting the "behavioral intelligence" that closely combines understanding and action. This embodied form of intelligence is more suitable for solving complex tasks in the open world, showing stronger real-world adaptability and application potential.

2.2. Analysis of difficulties of embodied intelligence

The core elements of EI and its intrinsic connection cause difficulties in its development, and there are three main breakthrough difficulties: building a powerful universal ontology platform, designing an intelligent agent system with advanced cognitive capabilities, and collecting high-quality industry data.

2.2.1. Building a powerful universal ontology platform

To achieve EI, we first need an ontology platform with excellent movement and operation capabilities. However, breaking through the bottleneck of key hardware technology and building robot products with high reliability, controllable costs, and strong versatility is still a very challenging task. In the context of pursuing "universal capabilities," humanoid robots are considered to be one of the final forms of EI. Therefore, research and development around humanoid structures will continue to be a hot topic and core problem that the industry and academia are concerned about [28].

2.2.2. Designing intelligent agent systems with advanced cognition

As the core element of EI, intelligent agent systems must not only adapt to the complex and ever-changing real environment but also have a series of advanced cognitive and interactive capabilities [29]. Specifically, intelligent agents need to solve the following key problems:

1) The ability to accurately perceive the three-dimensional physical environment;
2) Task scheduling, execution, and dynamic adjustment capabilities;
3) Strong general knowledge and multi-level semantic reasoning capabilities;
4) Natural and fluent human–computer language interaction (especially multi-round dialogue);
5) Construction and call of long-term memory mechanism;
6) Personalization and emotional care capabilities;
7) Cross-task generalization capabilities and self-learning and migration capabilities.

In order to adapt to the complexity of the real environment, embodied intelligent agents must have real-time perception and decision-making capabilities, which puts extremely high demands on the data collection, transmission, and processing speed of the system. At the same time, the current mainstream LLM has huge computational overhead. For resource-constrained robot embedded systems, it is a very challenging task to complete high-complexity reasoning and decision-making while ensuring low-latency response.

2.2.3. Lack of high-quality industry data is a bottleneck

In the real world, scenarios are highly complex and dynamically changing. Currently, there is a lack of sufficiently rich, diverse, and high-quality real-scene data to train a truly "universal" large model. Especially in critical businesses, the requirements for task success rates are extremely high, and it is difficult to meet actual application needs by relying solely on wide-area data.

A major feature of EI is that it is highly coupled with the physical environment, and effective data can only be obtained in real deployments, which is completely different from non-embodied intelligent systems that rely on pre-collected data. Therefore, in key areas, it is particularly important to obtain and build high-quality vertical industry data sets [30]. At the same time, by designing the intelligent agent structure in a hierarchical manner and limiting different tasks to specific scenarios, it is a feasible strategy to strike a balance between improving generalization ability and task success rate.

2.2.4. Integrate virtual and real interactions to achieve continuous evolution

The key to the evolution of EI lies in its ability to continuously learn and self-improve. In order to adapt to environmental changes, the agent must be able to actively learn new tasks and continuously optimize strategies during execution. An agent that adapts to its morphological characteristics can often master problem-solving methods faster and thus better adapt to new environments.

However, since the space of agent morphological design is almost infinite, it is not feasible to exhaust all possibilities under limited computing resources. In addition, the degree of freedom of the ontology design will also impose physical limitations on its task adaptability and learning ability, thereby affecting the learning and decision-making effects of the controller [31].

Therefore, there is a deep implicit relationship between complex environments, agent morphological evolution, and task learning ability that has not yet been fully revealed. How to achieve fast and efficient strategy learning and reasonable decision-making under limited resources will become a key breakthrough point for the future development of EI.

2.3. The latest breakthroughs in embodied intelligence

EI is an emerging interdisciplinary field that integrates physical embodiment with cognitive processes, enabling systems to engage in meaningful interaction with their environments. This literature review synthesizes recent contributions to the conceptual understanding and practical application of EI across domains such as robotics, soft actuators, and adaptive systems.

The latest breakthroughs in EI are shown in Table 2. Ma et al. [32] introduce an innovative design for soft actuators that integrates sensing, actuation, and control within individual units. Their research demonstrates the feasibility of creating autonomous soft robotic systems that exhibit EI, offering a streamlined yet effective methodology that advances the functionality and autonomy of soft robots. Gupta et al. [33] present deep evolutionary reinforcement learning (DERL), a computational framework that evolves agent morphologies to address locomotion and manipulation tasks in complex environments. Their findings reveal strong correlations between environmental complexity, morphological intelligence, and the learnability of control strategies, underscoring the adaptive potential of EI in dynamic and unpredictable contexts. They argue that EI should not be treated merely as an application area for machine learning, but rather as a driving force for its advancement. This perspective encourages a reevaluation of how embodied systems can inform and enhance the development of learning algorithms. Mengaldo et al. [34] further investigate physical modeling in soft robotics by providing a concise guide to the underlying physics of EI. Their work highlights the importance of grounding system design in physical principles to

Table 2
Comparison table of embodied intelligence-related research

Author	Year	Research topic	Key contribution	Application domain
Ma et al. [32]	2021	Integration of sensing and control in soft robotics	Designed soft actuators with embedded sensing, actuation, and control	Autonomous soft robotic systems
Gupta et al. [33]	2021	Evolutionary learning and morphological intelligence	DERL framework reveals correlation between environmental complexity and adaptive morphology	Robotic locomotion and manipulation
Mengaldo et al. [34]	2022	Physical modeling in soft robotics	Provided concise design guidelines based on physical principles	Adaptive robotic systems
Iida and Giardina [35]	2023	Temporal dimensions of Embodied Intelligence	Explained how self-organization emerges across different timescales	Autonomous adaptive systems
Fan et al. [36]	2025	Survey on LLMs in robotic systems	Emphasized the role of LLMs in robot autonomy and interaction	Intelligent robotic systems
Suo et al. [37]	2025	EI applications in sports science	Explored integration of digital human models with EI	Sports performance optimization and training

develop more responsive and adaptive robotic systems capable of real-world interaction. Iida and Giardina [35] contribute to the discourse by examining the temporal dynamics of EI, particularly within autonomous adaptive systems. Their structured review elucidates how varying timescales influence self-organization and the emergence of complex behaviors, offering insights into the temporal dimensions that underpin embodied functionality. The integration of LLMs into robotics is explored by Fan et al. [36], who examine the potential of LLMs to enhance robot intelligence, autonomy, and human–robot interaction. Their surveys underscore the transformative impact of LLMs on robotic perception, control, decision-making, and path planning, marking a significant step toward more versatile and intelligent robotic agents. Finally, Suo et al. [37] explore the convergence of digital human models and EI in the context of sports science. Their work identifies emerging trends and research opportunities, demonstrating how principles of EI can be applied to optimize performance and training methodologies across athletic disciplines.

As presented in Table 2 above, the reviewed studies collectively highlight the multidisciplinary progress in EI, spanning co-design frameworks, soft robotics, evolutionary learning, and LLM integration. Key contributions include structured modeling approaches, integrated soft actuators, adaptive morphology through DERL, and theoretical insights into EI-driven machine learning. Further developments explore physical modeling, temporal dynamics, human-in-the-loop systems, LLM-enhanced robotics, and applications in sports science, underscoring EI's potential across robotics, interaction, and intelligent system design.

3. Capabilities and Limitations of LLM

The core capabilities of LLM include linguistic generation, knowledge generalization, in-context learning, semantic comprehension, and multimodal Extension. However, LLM has many shortcomings, and the main limitations are context window restriction, illusion problem, knowledge limitation and lag, lack of comprehension, and poor interpretability. Specific explanations of the above issues are provided below.

3.1. Origin of LLM

LLM, the full name of large language model, is a large language model [38]. LLM is a powerful AI algorithm that can model natural language text by training a large amount of text data to learn the grammar, semantics, and contextual information of the language. This model has a wide range of applications in the field of NLP, including text generation, text classification, machine translation, sentiment analysis, etc. This article will introduce in detail the principles, development history, training methods, application scenarios, and future trends of the LLM large language model [39].

The evolution of the LLM can be roughly divided into three key development stages: statistical machine translation (SMT) stage, deep learning stage, and pre-training model stage.

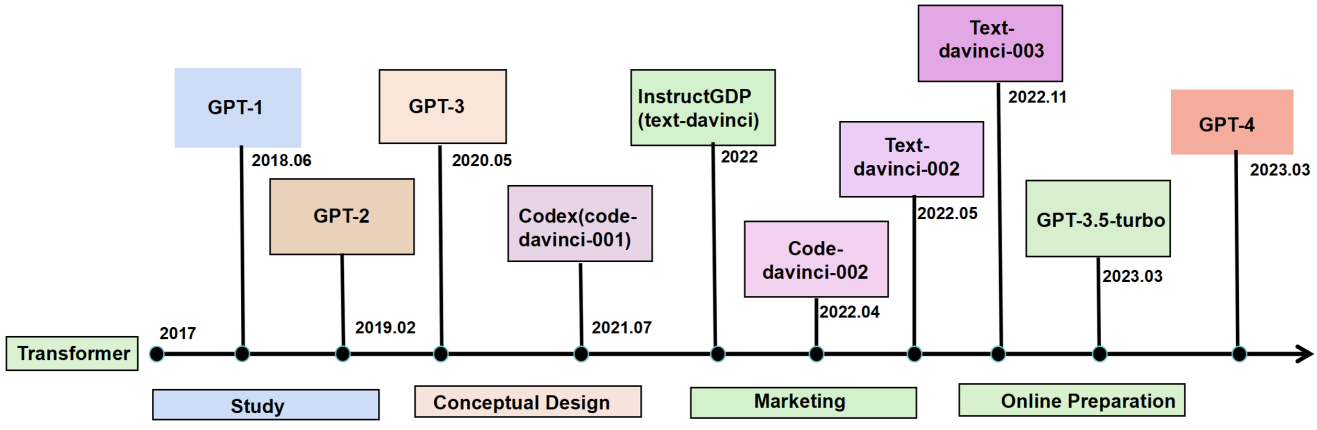
The first stage is the SMT period, which began around the beginning of the 21st century. In this stage, statistical methods became the mainstream NLP technology [40]. This method relies on large-scale bilingual corpora and uses statistical means to model the alignment relationship between the source language and the target language. Although SMT made important breakthroughs at the time, its performance in dealing with complex grammatical structures and long sentence translation was relatively limited, and it was difficult to capture deep semantic relationships.

The second stage began with the rise of deep learning technology. In 2013, the introduction of the word2vec model marked the rapid development of word vector technology. This technology maps discrete words into continuous low-dimensional vectors, effectively capturing the semantic associations between words. On this basis, neural structures such as recurrent neural networks (RNN), long short-term memory networks, and gated recurrent units have been introduced into NLP tasks, significantly improving the ability of language modeling and sequence generation.

The third stage entered the era of large-scale language models based on pre-training [41]. In 2018, the BERT model launched by Google became a landmark achievement. For the first time, the bidirectional Transformer architecture was adopted on a large scale to learn the deep semantic representation of language from massive, unlabeled text. The success of BERT triggered extensive research and application of the Transformer model, followed by a series of high-performance pre-trained language models, such as the GPT series, RoBERTa, and XLNet. These models have achieved unprecedented performance breakthroughs in various NLP tasks, promoting a paradigm shift in the entire field.

Figure 3 shows the development timeline of the GPT series of language models. From the birth of the Transformer architecture in 2017, to the start of the research phase of GPT-1 in 2018, followed by the continuous evolution of GPT-2 and GPT-3, during which variants such as Codex were derived. InstructGPT, text-davinci series, GPT-3.5-turbo, and GPT-4 were launched in 2022–2023,

Figure 3
LLM development origin diagram



going through stages from concept design to online preparation, showing their continuous iterative upgrade process.

After entering 2020, with the advent of super-large-scale models such as GPT-3, the development of LLMs has entered a period of rapid expansion. The scale of the model continues to grow, and the number of parameters has jumped from hundreds of millions to hundreds of billions, driving the implementation of many application scenarios such as AI writing, automatic question and answer, code generation, and multimodal interaction [42]. At the same time, major technology companies around the world have accelerated their layout to promote the rapid evolution of large models from academic research to industrialization and ecology. In September 2020, OpenAI authorized the GPT-3 model it developed to Microsoft for use, making Microsoft the first company to obtain the right to use GPT-3 [43]. Subsequently, OpenAI launched the public natural language generation model ChatGPT in 2022. On March 15, 2023, its multimodal upgraded version GPT-4 was officially released, further expanding the capabilities of large models in image and text understanding and interaction. At the same time, Google has also accelerated its layout in the field of generative AI. In February 2023, its press conference debuted Bard, a dialogue system driven by the LaMDA LLM. Then, on March 22, Google announced that Bard had entered the public beta stage, initially open to users in the United States and the United Kingdom, and then gradually promoted to the global market [44].

Domestically, Baidu announced on February 7, 2023, that it would launch an LLM product "Wenxin Yiyan," which was officially launched on March 16. This product is based on Baidu's self-developed Wenxin large model, relying on Baidu Smart Cloud to provide Application Programming Interface and computing services to enterprises and institutions, aiming to build an industrial-level AI ecosystem and promote the application of large model technology in various business scenarios [45].

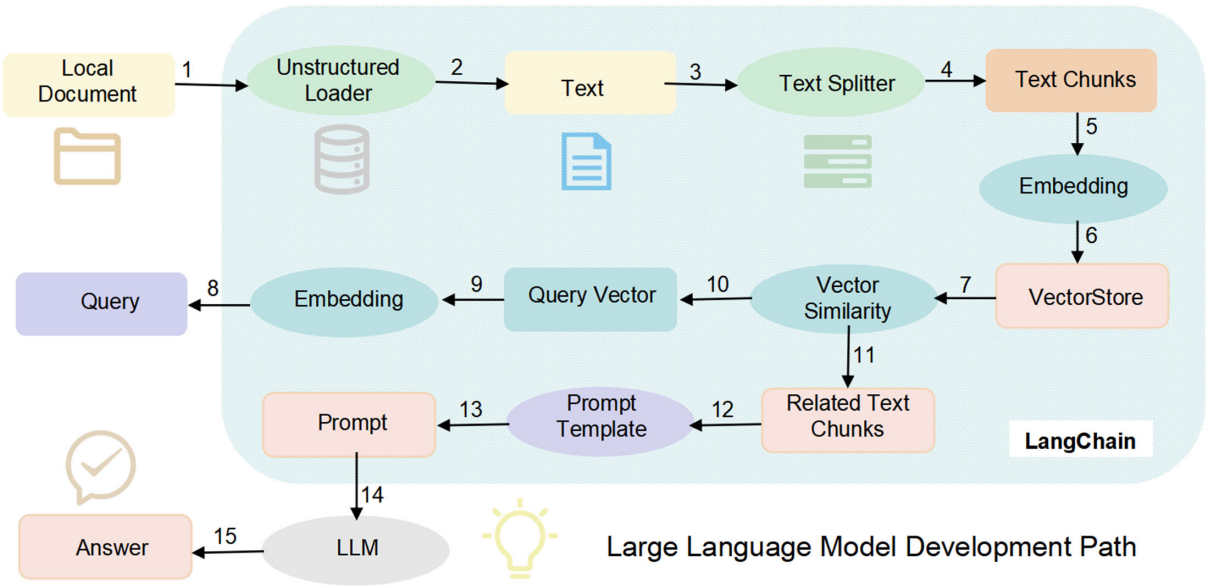
Amazon also announced its entry into the generative AI track on April 13, 2023, and launched a generative AI service called "Bedrock" through its cloud computing platform AWS. At the same time, it announced its self-developed LLM Titan, marking a key step in the field of enterprise-level AI services. Entering 2024, the field of open-source large models has also made important progress. In March 2024, Databricks released its latest open-source LLM DBRX, calling it "the strongest open source AI model at present," showing leading performance on multiple evaluation benchmarks [46].

In addition, in order to promote the standardized development of generative AI, in April 2024, at the 27th United Nations Science and Technology Conference held in Switzerland, the World Digital Technology Academy jointly released two international standards with OpenAI, Ant Group, iFlytek, Google, Microsoft, NVIDIA, Baidu, Tencent and other companies and research institutions: "Generative Artificial Intelligence Application Security Testing Standard" and "Large Language Model Security Testing Method." This marks an important step in the construction of the global governance and security system of LLMs.

3.2. Introduction to LLM underlying technology

The underlying technology of the LLM incorporates a number of core AI and deep learning approaches [47]. It is based on deep neural networks, in particular the model structure represented by the Transformer architecture, with a self-attention mechanism that enables powerful context modeling capabilities [48]. At the training level, strategies such as pre-training fine-tuning paradigms, language modeling objectives (e.g., autoregressive and masked language modeling), optimization algorithms (e.g., Adam), regularization, and so on are employed to improve model effectiveness and generalization. To cope with large-scale parameters and data, efficient computational techniques such as distributed training, mixed-precision computation, model parallelism, and data parallelism are used. Word embedding and positional coding techniques help the model to understand text semantics and order, while for model alignment, reinforcement learning and human feedback (e.g., RLHF) enhance the model's responsiveness to human intent. Together, these underlying technologies support the current strong performance of LLM in natural language understanding and generation. Figure 4 below depicts the development path of an LLM.

Figure 4
LLM technical composition structure diagram



First, the local document is processed into text by the unstructured loader, then split into text blocks by the text segmenter, and then the embedded vector is generated and stored in the vector storage. When querying, the query is converted into a vector, and the relevant text blocks are retrieved by vector similarity. The prompt template is combined to form a prompt, and the LLM is input to finally get the answer. Each link is closely connected to build the development process.

3.2.1. Neural network basic technology

Neural networks, also known as artificial neural networks (ANN) or simulated neural networks, are a key subfield of machine learning and a core pillar of deep learning algorithms. They are called "neural networks" because their structure and operation mechanism are inspired by the biological neural network in the human brain. Just as many technological inventions are derived from imitating nature (such as aircraft design is derived from bird flight), neural networks also imitate the signal transmission mechanism between neurons in the human brain to build intelligent systems that can learn and generalize autonomously [49].

In the formal definition, an ANN is a computational model composed of multiple "neurons" or nodes, which are connected to each other in the form of layers. Each neuron processes the input with weights, calculates the weighted sum, and then applies a nonlinear activation function to pass the information layer by layer to finally obtain the model output. Such a structure can abstract and extract high-order features from the data layer by layer and is a powerful tool for identifying complex patterns and relationships.

The concept of neural networks can be traced back to 1943, when Warren McCulloch and Walter Pitts proposed the first mathematical model to simulate the behavior of neurons in the human brain, which is also regarded as the first ANN. Later in the 1950s, Frank Rosenblatt designed the "Perceptron" model, a simple two-layer neural network that can be trained to recognize linearly separable patterns. Although the perceptron attracted great attention at the time, its limitations in dealing with nonlinear problems caused research to stagnate. It was not until the 1980s that Geoffrey Hinton and others proposed the "backpropagation algorithm," a breakthrough that enabled neural networks to effectively train multi-layer structures and thus learn more complex nonlinear mapping relationships. The introduction of backpropagation opened a new chapter in neural networks and laid the foundation for later deep learning [50].

In the 1990s, although deep learning research was still on the academic fringe, related explorations continued to advance. The real turning point came in the early 2000s, with the rise of big data and the leap-forward improvement of computing resources, deep neural networks achieved performance breakthroughs in many fields. The most representative achievements include the widespread application of convolutional neural networks (CNN) and RNN. CNN is widely used in tasks such as image recognition, face recognition, and object detection, while RNN performs well in tasks that process sequence data such as speech recognition, NLP, and machine translation [51].

With the improvement of algorithms, the increase of computing power, and the availability of large-scale training data, deep learning has made significant progress in many sub-fields of AI in the past decade, especially in tasks such as image recognition, speech recognition, natural language understanding, and generation, showing the ability to surpass traditional machine learning methods. Today, neural networks and their deep learning variants have become one of the core technologies driving the development of AI, and one of the most active and promising research directions at present, providing a solid technical foundation for advanced AI systems including LLMs.

3.2.2. Deep learning core architecture: Transformer

As a major breakthrough in the field of deep learning, the Transformer model has become the core model structure in NLP and even the entire field of AI. Its innovative design completely abandons the traditional RNN and CNN structures and instead introduces a fully parallelized architecture based on the attention mechanism, which greatly improves the model's training efficiency and long-distance dependency modeling capabilities. The core of the Transformer lies in its encoder-decoder architecture, supplemented by key components such as positional encoding, multi-head attention mechanism, and feedforward neural network, which together constitute a powerful semantic modeling capability [52].

1) Encoder-decoder architecture

The basic structure of the Transformer is a typical encoder-decoder architecture, which consists of multiple stacked encoder and decoder layers. Each layer adopts a modular design to facilitate parallel computing and expansion.

The encoder is responsible for receiving the input sequence and converting it into a set of hidden representations containing rich semantic information. Each encoder layer contains two sublayers: multi-head self-attention mechanism and feedforward network, plus residual connection and layer normalization.

The decoder receives the output representation of the encoder in the generation phase and generates the next word based on the existing output context. The decoder layer contains three sublayers: self-attention mechanism, encoder-decoder attention mechanism, and feedforward network. This structure supports context-sensitive language generation and is widely used in tasks such as machine translation, text summarization, and dialogue systems.

2) Positional encoding

Since Transformer is completely based on the attention mechanism and lacks the inherent ability to model sequence structure, it is necessary to explicitly introduce information about the sequence order – this is the role of positional encoding [53].

The sine and cosine function position encoding commonly used in the Transformer adds a set of distinguishable position features to each input vector, allowing the model to perceive the relative or absolute position relationship between words.

This encoding method has excellent generalizability and can effectively migrate to new sequence lengths that do not appear in the training data.

3) Multi-head attention

Multi-head attention is one of the most representative technical innovations in the Transformer. It allows the model to learn information from multiple subspaces simultaneously, enhancing the diversity and expressiveness of representation.

Each "head" independently learns a set of attention weights from the input to capture semantic relationships of different dimensions and levels. For example, one attention head may focus on the subject-predicate structure, another on the emotional color, and the third on the entity relationship [54].

In terms of operation, the input query, key, and value vectors are linearly mapped to multiple subspaces, respectively, and the attention operation is performed in parallel in these subspaces. Finally, the outputs of all attention heads are spliced and linearly transformed to summarize the final result.

The Transformer model has set a new technical benchmark in the field of NLP with its highly modular architecture and innovative mechanisms [55]. From the bidirectional modeling capabilities provided by the encoder-decoder architecture, to the sequence awareness brought by positional encoding, to the powerful representation capabilities of multi-head attention and feedforward networks, each component complements each other and together gives the Transformer powerful language understanding and generation capabilities. It is the synergy of these underlying mechanisms that has enabled LLMs represented by GPT, BERT, T5, etc., to develop rapidly and be widely used in multiple AI application scenarios such as search, question and answer, translation, dialogue systems, content generation, etc. [56].

3.2.3. Reinforcement learning and alignment techniques

In the field of multimodal learning, different modal data have very different statistical properties and representations, and the abstraction level of the same concept in different modalities is not consistent, making dynamic interaction difficult to realize. Multimodal alignment technology is committed to establishing structured associations of different modal data in a unified semantic space and bridging the semantic gap between modalities through representation learning, attention mechanism, and generative modeling. The core lies in constructing cross-modal mapping functions so that the feature vectors of heterogeneous data satisfy semantic consistency in the embedding space: for example, relevant graphic pairs are mapped to neighboring regions, while irrelevant samples are separated from each other. To achieve this goal, mainstream approaches rely on three main paradigms: first, contrastive representation learning, for example, CLIP, ALIGN models, which utilize large-scale noisy modal pair data to implicitly match global semantics via contrastive loss functions; second, cross-modal attention mechanisms, for example, ViLBERT, Flamingo, which dynamically model inter-modal interaction dependencies on local features via gated attention weights such as the association of textual lexical items with image regions; and third, generative alignment, for example, DALL-E, BLIP, which realizes bidirectional semantic mapping with the help of explicit constraints on joint distributional learning of cross-modal generative tasks.

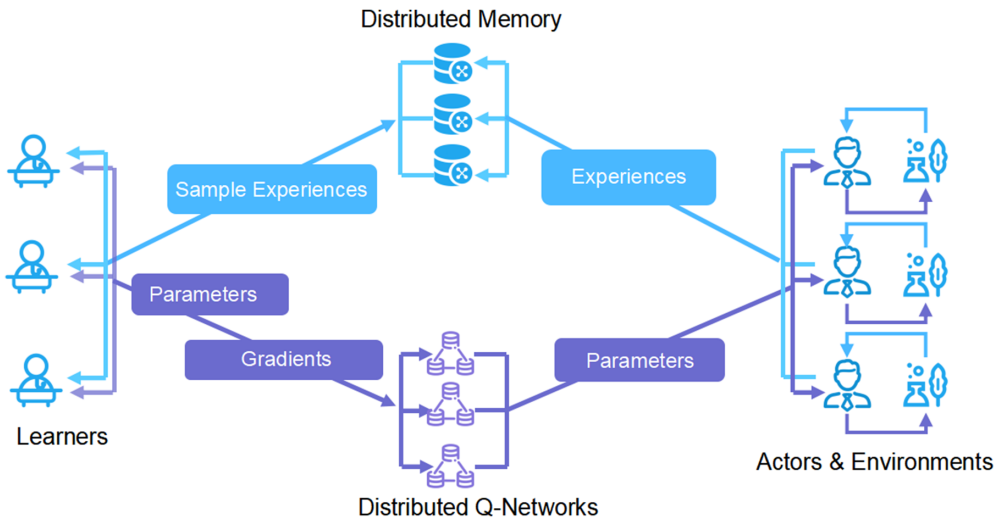
There are three main paradigms for the alignment of LLMs: supervised fine-tuning-based alignment, reinforcement learning-based alignment, and in-context learning (ICL). These paradigms use different mechanisms to make the model more in line with human expectations and values. Among them, supervised fine-tuning relies on high-quality annotated data for model adjustment, reinforcement learning guides model behavior through reward signals, and ICL controls output behavior through contextual cues without changing model parameters, gradually becoming a cost-effective alternative. Figure 5 illustrates the distributed reinforcement learning architecture [57].

As can be seen from Figure 5, learners acquire sample experience, and parameters and gradients are updated in the distributed Q network. Experience is stored in distributed memory for actors to use in different environments. The interaction between actors and the environment generates new experience, forming a cycle. This distributed architecture improves the efficiency and effectiveness of reinforcement learning and handles complex tasks [58].

The application of reinforcement learning, especially in RLHF (reinforcement learning with human feedback), has greatly promoted the alignment of LLMs with human intentions. RLHF was first proposed in 2017 to improve the behavioral complexity of deep reinforcement learning models through human preference feedback on agent behavior fragments. Subsequently, RLHF was introduced to tasks such as text summarization, significantly improving the output quality. In WebGPT, this technology is used to optimize web page retrieval and information aggregation capabilities. Although these early applications mainly focus on improving the "usefulness" and "honesty" of the model, "harmlessness" is often overlooked, which may lead to model outputs that violate human values.

To make up for the shortcomings of RLHF, InstructGPT annotates model responses through human feedback, thereby achieving deep alignment with user intent and meeting the standards of "usefulness, honesty, and harmlessness (HHH)." The success of this technology directly gave birth to ChatGPT, which has become one of the most successful interactive LLMs to date and has injected new vitality into the research of artificial general intelligence (AGI).

Figure 5
Reinforcement learning process schematic diagram



However, RLHF has significant challenges in training: it usually requires running multiple LLMs at the same time and relies on a large amount of high-quality manually annotated data, resulting in high data and training costs. To this end, researchers proposed the Constitutional AI framework to reduce dependence on manual labels [59]. This method uses LLM to generate and correct its own responses, replaces manual feedback with AI feedback, and thus forms a new path for reinforcement learning based on AI feedback (RLAIF). An important variant of RLAIF is to train the reward model using synthetic preference data from different model scales and prompt conditions, then automatically generate demonstration data for supervised fine-tuning, and finally fine-tune the LLM policy through reinforcement learning. In order to further improve the training efficiency, the ReST (Reinforcement Self-Training) method is also proposed to expand the training data by sampling responses from existing policies and update the model policy under offline Reinforcement Learning objectives [60].

Although RLHF has good generalization ability and shows great potential for leveraging human feedback signals, the instability of its training process and high cost still restrict its application in a wider range of scenarios. In addition, the trade-off between different objectives and the lack of a normative loading mechanism are still challenges in achieving robust alignment.

3.3. LLM latest research

LLMs have attracted considerable scholarly interest in recent years due to their transformative potential across a wide range of applications, spanning NLP, personalized learning, and multimodal systems [61]. This literature review synthesizes recent advancements and methodological innovations related to LLMs, with a focus on training paradigms, personalization techniques, and application strategies. Table 3 describes and compares the results of research related to LLM.

In the context of language learning, Wang and Reynolds [62] investigate how Chinese learners of English engage with LLMs for vocabulary acquisition. Their study identifies effort expectancy as a key determinant of users' intention to adopt LLMs and highlights the underexplored potential of these models in informal learning environments. The findings suggest that LLMs could serve as effective tools in nontraditional educational settings, offering personalized and adaptive support. Yao and Yuan [63] provide a comprehensive overview of language representation techniques, model architectures, and contemporary applications of LLMs. Their work emphasizes the need for systematic exploration of optimization strategies to enhance model performance and scalability. This foundational understanding is critical for advancing the practical deployment of LLMs in diverse domains. Addressing computational challenges in model distillation, Ko et al. [64] present DistiLLM, a framework designed to resolve training-inference mismatches in autoregressive LLMs. Their work demonstrates the importance of efficient distillation techniques for improving inference efficiency and maintaining performance in large-scale systems. Personalization remains a pivotal area in LLM research. Salemi et al. [65] explore retrieval-augmented generation as a means of tailoring LLM outputs, optimizing retrieval models that supply user-relevant documents to enhance the relevance and coherence of generated text. This research is further extended in their subsequent study on LaMP, which systematically compares retrieval strategies for adapting language model behavior to individual user profiles. Zheng et al. [66] introduce SGLang, a system that enables efficient execution of structured LLM programs. By supporting advanced prompting and control flow mechanisms, SGLang facilitates complex task execution and expands the functional capabilities of LLMs in real-world applications.

Finally, the integration of vision and language is examined by Diao et al. [67], who develop a training methodology for encoder-free vision-language models (VLMs). Their work bridges the gap between encoder-based and encoder-free paradigms, contributing to the development of more flexible and efficient multimodal systems.

Recent advancements in LLMs span multiple fronts, including training methods, personalization, optimization, and multimodal integration. Yao and Yuan [68] emphasize the need for systematic optimization. Ko et al. [69] introduce DistiLLM to resolve training-inference mismatches. Personalization strategies are advanced by Salemi et al. [70] via retrieval-augmented generation.

Table 3
Comparison table of LLM-related research

Authors	Research topic	Method/model	Key findings and contributions
Wang et al. [62]	Vocabulary acquisition by English learners using LLMs	Empirical study; surveys and behavioral data	Identifies effort expectancy as a key factor for adoption; highlights the potential of LLMs in informal learning contexts.
Yao et al. [63]	Review of LLM optimization strategies	Review; language representation, architecture, applications	Stresses the importance of systematic exploration of optimization strategies for broader deployment.
Ko et al. [64]	Addressing training-inference mismatch via distillation	DistiLLM framework; autoregressive models	Mitigates inconsistencies between training and inference, improving efficiency in large-scale models.
Salemi et al. [65]	Retrieval-augmented generation for personalization	Personalized retrieval models; LaMP framework	Improves relevance and coherence of generated outputs by adapting to user-specific context and profiles.
Zheng et al. [66]	Efficient execution of structured LLM programs	SGLang system; structured prompting, control flow	Supports complex task execution and expands programming flexibility in real-world applications.
Diao et al. [67]	Training of encoder-free vision-language models	Encoder-free training for multimodal systems	Bridges encoder-based and encoder-free approaches, contributing to more flexible and efficient multimodal LLMs.

Zheng et al. [71] present SGLang for structured LLM execution, and Diao et al. [72] bridge encoder-based and encoder-free VLMs, supporting more flexible multimodal systems.

4. The Integration Path of Embodied Intelligence and LLM

The advancements in LLMs have demonstrated their remarkable capabilities in knowledge acquisition and reasoning. The integration of EI with LLMs has the potential to endow intelligent systems with both language comprehension and sensory perception, thereby significantly enhancing their reasoning abilities and action execution. In this context, EI emerges as an autonomous agent capable of interacting with the environment and making decisions through self-directed planning. Ongoing research efforts are actively exploring this integration, and notable progress has already been achieved. Figure 6 summarizes the development of AI and robots. In the early days, AI simulated human thinking with symbolism and expert systems and then shifted to data-driven machine learning and deep learning; EI started from automation and industrial applications and faced bottlenecks such as mobile collaboration capabilities and adaptability to complex environments. In the future, the two need to be integrated and developed to achieve the unification of perception and interaction and learning and adaptation and build "embodied" AI with powerful intelligence.

Gürcan et al. [68] propose an LLM-augmented world model that enhances embodied agents' planning abilities through visual perception and prediction-oriented prompts. Implemented in the Minecraft environment using the VOYAGER framework, the system enables agents to autonomously explore, plan, and complete tasks. The study demonstrates that integrating visual data and explicitly guiding LLM predictions significantly improves performance. However, the model is limited by its dependence on specific LLMs and constrained task scope, highlighting the need for broader applicability and validation.

Yang et al. [69] is a language-driven environment generation system that creates interactive 3D scenes from natural language prompts. Using GPT-4 and a large asset library, it generates realistic room layouts and populates them with physically grounded, interactable objects. The system is integrated with the AI2-THOR simulator to enable embodied AI agents to train and evaluate in diverse, customizable virtual spaces. Limitations include reliance on predefined assets and indoor-only environments, with future plans to support broader scene categories. Peng et al. [70] introduce a human-like AI interviewer, implemented using the android ERICA. The system demonstrates adaptive conversational behavior including listening, repairing dialogue, and providing post-interview summaries via LLM-based processing. Deployed at a real-world conference, the system received positive human feedback. Limitations include the use of templated questions, lack of multimodal input, and limited participant diversity. Dai et al. [71] explore the use of AI-assisted flexible electronics to create authentic and expressive facial gestures in humanoid robots. It reviews developments in biomimetic facial design, including sensors, actuators, and intelligent artificial skin capable of thermal and color modulation. Applications span social, companion, and service robotics. Key challenges include integrating perception and actuation, emotional reasoning, and the need for interdisciplinary materials and AI collaboration.

PHYSCENE is a 3D scene generation system designed for embodied AI training with physical interactivity in mind. It uses conditional diffusion guided by constraints such as object reachability, layout rules, and collision avoidance. The resulting scenes are suitable for manipulation and navigation tasks. Limitations include support for only certain room types and a lack of small object interaction capabilities [72]. PR2 is a physics- and photo-realistic humanoid robot simulation platform designed for education, competition, and research. It supports the integration of foundation models, advanced planning algorithms, and dynamic interaction tasks. The platform was piloted in a national competition and facilitates both locomotion and manipulation learning. Challenges include sim-to-real gaps and continued development for broader task compatibility.

As demonstrated in Table 4, the integration of EI and LLMs enhances the cognitive-executive feedback loop through multimodal perception, real-time reasoning, and dynamic planning. This integration facilitates a shift from passive responses to proactive decision-making, marking a critical pathway toward AGI. Bousetouane introduced a modular architecture comprising perception, cognition, and actuation components. Moreover, researchers such as Pang, Dai, and Liu have applied the fusion of EI and LLMs to real-world contexts, including AI interviewers, healthcare, and humanoid robot simulation platforms, significantly enhancing operational efficiency in these domains.

Figure 6
The development history of the combination of embodied intelligence and AI

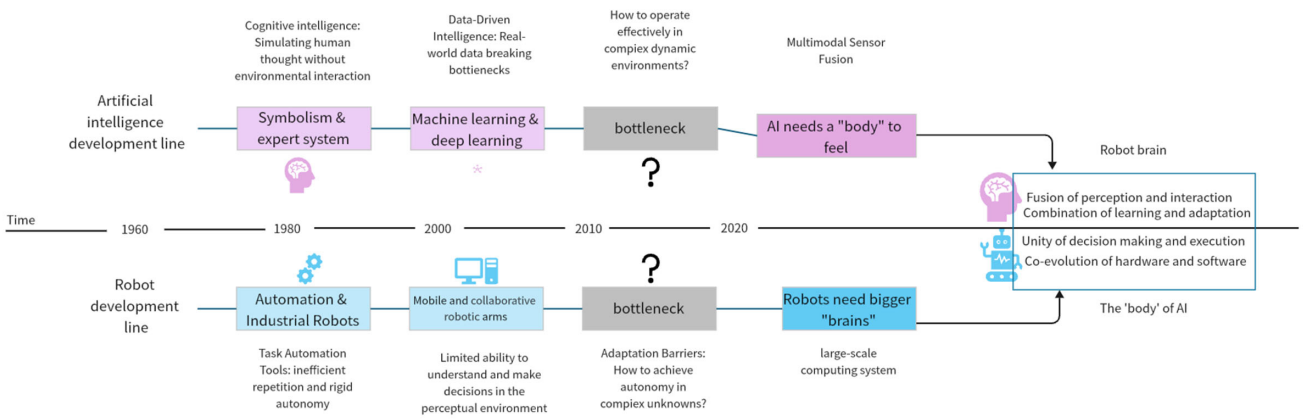


Table 4
Comparison table of embodied intelligence and LLM-related research

Authors	Innovation highlight	Techniques/methods	Application domain	Challenges/limitations
Gürcan et al. [68]	Embodied world model integrating LLM with visual perception	LLMs (GPT), visual encoding, prediction prompts	Simulation (Minecraft exploration tasks)	Specific LLM constraints, limited task generality
Yang et al. [69]	Language-driven automatic generation of interactive 3D environments	LLMs (GPT-4), constraint-based optimization, Objaverse 3D assets	Simulation (navigation, interaction tasks)	Limited asset types, mainly indoor scenarios
Peng et al. [70]	Human-like conversational interviewer android	Conversational AI, LLM-based analysis, adaptive dialogue	Human-robot interaction, qualitative interviewing	Small sample size, rigid question templates
Dai et al. [71]	Natural facial expressions via AI-assisted flexible electronics	Flexible sensors, artificial skin, multimodal sensing	Social robotics, interactive humanoid robots	Integration of sensing, actuation, emotional cognition

5. Future Directions

EI integrated with LLMs exhibits a high degree of interactivity and adaptability. By executing a perception–action loop, it can flexibly adjust its behavior in response to environmental changes. This capability relies on hardware components such as sensors and actuators, as well as sophisticated control algorithms and machine learning techniques. However, achieving further breakthroughs in EI requires addressing several critical challenges.

At present, EI is predominantly trained in virtual simulation environments, and there remains substantial room for progress in addressing the representation of complex multimodal information in real-world 3D settings. To meet the demands of practical applications, the multimodal unified models used in the integration of EI and LLMs are expected to evolve toward more comprehensive world modeling, greater autonomy, and hybrid large model architectures.

The integration of EI with LLMs is primarily driven by advancements in deep learning for robotics. End-to-end architectures are commonly adopted as system design paradigms, enabling efficient responses and emergent intelligence. However, these models require vast amounts of training data. Currently, the scale of robotic training data – typically ranging from thousands to tens of thousands of samples – falls significantly short of the billions required by VLMs, presenting a gap of several orders of magnitude. Moreover, the cost of acquiring high-quality data is substantial. While synthetic data generated in virtual environments offers an alternative, it often suffers from inaccurate physical modeling. The shortage of high-quality 3D datasets and the low level of data standardization further hinder progress. These challenges collectively pose significant obstacles to the advancement of EI. To address them, the development of open-source datasets and standardized benchmark environments is essential. However, due to significant differences in the physical design of robots produced by different manufacturers, open-source datasets often suffer from poor reusability and limited compatibility, further exacerbating the issue of data scarcity. To address this, cross-industry collaboration is essential to establish appropriate and standardized protocols that enhance the interoperability and reusability of training datasets.

In terms of benchmark environment construction, several high-quality platforms have been developed to support the training of EI. For example, Habitat, a high-performance 3D simulation platform developed by Facebook AI Research (FAIR), is designed to train embodied agents – such as robots – for navigation and interaction in complex virtual environments. It emphasizes learning and adaptation within simulated spaces that closely approximate real-world physical settings. ALFRED is a benchmark task framework that requires embodied agents to execute multistep tasks in a virtual environment based on natural language instructions. Its core lies in the integration of language understanding and physical interaction capabilities, combining visual, linguistic, and action data. iGibson is an interactive simulation environment that focuses on robotic physical interaction tasks in domestic settings. It features a high-fidelity physics engine and a diverse set of household scenes, enabling complex operations such as door opening and object manipulation. These platforms provide robust experimental foundations for the development of EI. However, due to the inherent complexity and variability of world models, it remains necessary to construct tailored benchmark environments for specific scenarios.

The advancement of EI integrated with LLMs relies heavily on collaboration between academia and industry. These two sectors are mutually reinforcing and indispensable, forming a relationship between technological complementarity and collaborative innovation. Cutting-edge developments in this field – such as the design of multimodal datasets and deep learning algorithms based on neural networks – require strong support from academic and research institutions. Meanwhile, the industrial sector focuses more on the development of embodied AI hardware and its application in real-world scenarios, translating technological innovations into practical use. This, in turn, provides feedback that helps steer academic research, creating a synergistic effect through the sharing of resources and knowledge. With continued interdisciplinary collaboration, the integration of EI and LLMs is poised for greater breakthroughs and transformative growth.

6. Conclusion

This paper provides and in-depth reviews of the relevant research on the integration of EI and LLMs. EI shows unique advantages by interacting with the environment, but it has shortcomings in semantic understanding and data acquisition. LLMs are excellent in language processing capabilities and can effectively make up for the shortcomings of EI. The integration of the two has significant potential and value, opening up a new path for the development of AI. With the continuous improvement and breakthrough of related technologies, it is expected to realize innovative applications in more fields and promote the development of AI in the direction of generalization and intelligence. Although this paper systematically reviews the research progress of the fusion of EI and LLMs, there are still some shortcomings that need further exploration. First, current research is mainly based on literature summary, lacking empirical experiments and quantitative evaluation, making it difficult to

fully verify the fusion effect; second, the summary of fusion paths is still relatively fragmented, lacking a unified system architecture model; at the same time, key issues such as data privacy, model interpretability, and ethical risks have not been deeply discussed; in addition, there is insufficient comparative analysis of existing open-source tool chains and standardized platforms, and the environmental complexity and reliability challenges faced in actual deployment are not fully covered. Future research should strengthen expansion in building a unified fusion paradigm, strengthening real-world scenario verification, and promoting cross-industry standard construction, so as to achieve the improvement of the versatility, security, and practicality of EI systems.

Funding Support

This work is sponsored by the Scientific Research Foundation of Chongqing University of Technology (0119240197). It was also supported by equipment funded through the "Intelligent Connected New Energy Vehicle Teaching System" project of Chongqing University of Technology, under the national initiative "Promote large-scale equipment renewals and trade-ins of consumer goods." Furthermore, this paper was supported by Innovative Research Group of Chongqing Municipal Education Commission (CXQT19026), and Cooperative Project between Chinese Academy of Sciences and University in Chongqing (HZ2021011).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Fujiang Yuan: Conceptualization, Methodology, Writing – original draft, Writing – original draft. **Hong Jiang:** Validation, Formal analysis, Resources, Writing – original draft, Writing – review & editing. **Haoran Guan:** Software, Investigation, Data curation, Visualization. **Yanhong Peng:** Supervision, Project administration, Funding acquisition.

References

[1] Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence*, *149*(1), 131–150. https://doi.org/10.1016/S0004-3702(03)00055-9

[2] Holland, O. (2004). The future of embodied artificial intelligence: Machine consciousness? In F. Iida, R. Pfeifer, L. Steels, & Y. Kuniyoshi (Eds.), *Embodied Artificial Intelligence: International seminar* (pp. 37–53). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-27833-7_3

[3] Manoonponga, P., Badri-Spröwitz, A., & Owaki, D. (2025). Special issue on embodied intelligence-understanding animal locomotion and its robotic implementations. *Advanced Robotics*, *39*(1), 1–2. https://doi.org/10.1080/01691864.2024.2448348

[4] He, Z., Duan, H., Zeng, J., Zhou, J., Zhong, X., Wu, Z., . . . , & Liu, G. (2025). Perovskite retinomorphic image sensor for embodied intelligent vision. *Science Advances*, *11*(1), eads2834. https://doi.org/10.1126/sciadv.ads2834

[5] Liu, J., Hao, W., Cheng, K., & Jin, D. (2025). Large language model-based planning agent with generative memory strengthens performance in textualized world. *Engineering Applications of Artificial Intelligence*, *148*, 110319. https://doi.org/10.1016/j.engappai.2025.110319

[6] Sun, Y., & Ji, F. (2025). An embodied intelligence system for coal mine safety assessment based on multi-level large language models. *Sensors*, *25*(2), 488. https://doi.org/10.3390/s25020488

[7] Sun, X., Mangan, M., Peng, J., & Yue, S. (2025). I2Bot: An open-source tool for multi-modal and embodied simulation of insect navigation. *Journal of The Royal Society Interface*, *22*(222), 20240586. https://doi.org/10.1098/rsif.2024.0586

[8] Wang, Y., Shi, M., Liu, J., Zhong, M., & Ran, R. (2025). The impact of digital-real integration on energy productivity under a multi-governance framework: The mediating role of AI and embodied technological progress. *Energy Economics*, *142*, 108167. https://doi.org/10.1016/j.eneco.2024.108167

[9] Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., . . . , & Chen, J. (2025). From screens to scenes: A survey of embodied AI in healthcare. *Information Fusion*, *119*, 103033. https://doi.org/10.1016/j.inffus.2025.103033

[10] Hassanin, M., Keshk, M., Salim, S., Alsubaie, M., & Sharma, D. (2025). PLLM-CS: Pre-trained Large Language Model (LLM) for cyber threat detection in satellite networks. *Ad Hoc Networks*, *166*, 103645. https://doi.org/10.1016/j.adhoc.2024.103645

[11] Wang, H. W., Hoffswell, J., Thazin Thane, S. M., Bursztyn, V. S., & Bearfield, C. X. (2025). How aligned are human chart takeaways and LLM predictions? A case study on bar charts with varying layouts. *IEEE Transactions on Visualization and Computer Graphics*, *31*(1), 536–546. https://doi.org/10.1109/TVCG.2024.3456378

[12] Zack, T. I., Sushil, M., Miao, B., Demirci, A., Kasap, C., Tempero, M., . . . , & Collisson, E. (2024). Abstract B074: Clinical inference of location and trajectory of pancreatic cancer from radiology reports using zero-shot LLM. *Cancer Research*, *84*(2_Supplement), B074–B074. https://doi.org/10.1158/1538-7445.PANCA2023-B074

[13] Arslan, B., Nuhoglu, C., Satici, M. O., & Altinbilek, E. (2025). Evaluating LLM-based generative AI tools in emergency triage: A comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. *The American Journal of Emergency Medicine*, *89*, 174–181. https://doi.org/10.1016/j.ajem.2024.12.024

[14] Wong, W., Salisbury, T., & Pyatibrat, S. (2025). Development and clinical performance of an LLM-generated application for assessment of HER2 in situ hybridization in breast cancer. *Laboratory Investigation*, *105*(3), 102474. https://doi.org/10.1016/j.labinv.2024.102474

[15] Zheng, L., Jiang, F., Gu, X., Li, Y., Wang, G., & Zhang, H. (2025). Teaching via LLM-enhanced simulations: Authenticity and barriers to suspension of disbelief. *The Internet and Higher Education*, *65*, 100990. https://doi.org/10.1016/j.iheduc.2024.100990

[16] Duan, J., Li, Z., Wang, X., Li, W., Bai, Q., & Nguyen, M. (2025). Intent-spectrum bottracker: Tackling LLM-based social media bots through an enhanced BotRGCN model with intention and

entropy measurement. In *Knowledge Management and Acquisition for Intelligent Systems: 20th Principle and Practice of Data and Knowledge Acquisition Workshop*, 55–67. https://doi.org/10.1007/978-981-96-0026-7_5

[17] Efstathiadis, G., Yadav, V., & Abbas, A. (2025). LLM-based speaker diarization correction: A generalizable approach. *Speech Communication*, *170*, 103224. https://doi.org/10.1016/j.specom.2025.103224

[18] Gregory, G., & Vito, L. (2024). ChatGPT: A canary in the coal mine or a parrot in the echo chamber? Detecting fraud with LLM: The case of FTX. *Finance Research Letters*, *70*, 106349. https://doi.org/10.1016/j.frl.2024.106349

[19] Yin, W. (2007). Yǔyán shìjièguān duōyuánlùn—Bā lún yǔyán de tǐyàn guān [The pluralism of linguistic world outlook: The 8th paper on linguistic embodiment based on embodied philosophy and CL]. *Journal of Chongqing University (Social Science Edition)*, *13*(1), 112–117.

[20] Piran, N. (2016). Embodied possibilities and disruptions: The emergence of the Experience of Embodiment construct from qualitative studies with girls and women. *Body Image*, *18*, 43–60. https://doi.org/10.1016/j.bodyim.2016.04.007

[21] Paradowski, M. B. (2011). Developing embodied multisensory dialogue agents. *arXiv Preprint: 1111.7190*.

[22] Sandini, G., Sciutti, A., & Morasso, P. (2024). Artificial cognition vs. Artificial intelligence for next-generation autonomous robotic agents. *Frontiers in Computational Neuroscience*, *18*, 1349408. https://doi.org/10.3389/fncom.2024.1349408

[23] Agassounon, W., Martinoli, A., & Easton, K. (2004). Macroscopic modeling of aggregation experiments using embodied agents in teams of constant and time-varying sizes. *Autonomous Robots*, *17*(2–3), 163–192. https://doi.org/10.1023/B:AURO.0000033971.75494.c8

[24] Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*(8), 1503–1514. https://doi.org/10.1016/j.paid.2006.10.023

[25] Brannigan, G. (1975). Scoring difficulties on the Wechsler intelligence scales gary. *Psychology in the Schools*, *12*(3), 313–314. https://doi.org/10.1002/1520-6807(197507)12:3<313::AID-PITS2310120313>3.0.CO;2-K

[26] Lefebvre, S. (2003). The difficulties and dilemmas of international intelligence cooperation. *International Journal of Intelligence and CounterIntelligence*, *16*(4), 527–542. https://doi.org/10.1080/716100467

[27] Brankaer, C., Ghesquière, P., & de Smedt, B. (2014). Numerical magnitude processing deficits in children with mathematical difficulties are independent of intelligence. *Research in Developmental Disabilities*, *35*(11), 2603–2613. https://doi.org/10.1016/j.ridd.2014.06.022

[28] Mao, Z., Bai, X., Peng, Y., & Shen, Y. (2024). Design, modeling, and characteristics of ring-shaped robot actuated by functional fluid. *Journal of Intelligent Material Systems and Structures*, *35*(19), 1459–1470. https://doi.org/10.1177/1045389X241276216

[29] Peng, Y., Wang, Y., Hu, F., He, M., Mao, Z., Huang, X., & Ding, J. (2024). Predictive modeling of flexible EHD pumps using Kolmogorov–Arnold Networks. *Biomimetic Intelligence and Robotics*, *4*(4), 100184. https://doi.org/10.1016/j.birob.2024.100184

[30] Zhang, C., Chen, J., Li, J., Peng, Y., & Mao, Z. (2023). Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, *3*(4), 100131. https://doi.org/10.1016/j.birob.2023.100131

[31] Bai, X., Peng, Y., Li, D., Liu, Z., & Mao, Z. (2024). Novel soft robotic finger model driven by electrohydrodynamic (EHD) pump. *Journal of Zhejiang University-SCIENCE A*, *25*(7), 596–604. https://doi.org/10.1631/jzus.A2300479

[32] Ma, B., Xu, C., Cui, L., Zhao, C., & Liu, H. (2021). Magnetic printing of liquid metal for perceptive soft actuators with embodied intelligence. *ACS Applied Materials & Interfaces*, *13*(4), 5574–5582. https://doi.org/10.1021/acsami.0c20418

[33] Gupta, A., Savarese, S., Ganguli, S., & Fei-Fei, L. (2021). Embodied intelligence via learning and evolution. *Nature Communications*, *12*(1), 5721. https://doi.org/10.1038/s41467-021-25874-z

[34] Mengaldo, G., Renda, F., Brunton, S. L., Bächer, M., Calisti, M., Duriez, C., . . . , & Laschi, C. (2022). A concise guide to modelling the physics of embodied intelligence in soft robotics. *Nature Reviews Physics*, *4*(9), 595–610. https://doi.org/10.1038/s42254-022-00481-z

[35] Iida, F., & Giardina, F. (2023). On the timescales of embodied intelligence for autonomous adaptive systems. *Annual Review of Control, Robotics, and Autonomous Systems*, *6*(1), 95–122. https://doi.org/10.1146/annurev-control-063022-094301

[36] Fan, H., Liu, X., Fuh, J. Y. H., Lu, W. F., & Li, B. (2025). Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. *Journal of Intelligent Manufacturing*, *36*(2), 1141–1157. https://doi.org/10.1007/s10845-023-02294-y

[37] Suo, X., Tang, W., Mao, L., & Li, Z. (2025). Digital human and embodied intelligence for sports science: Advancements, opportunities and prospects. *The Visual Computer*, *41*(4), 2477–2493. https://doi.org/10.1007/s00371-024-03547-4

[38] Muhammad, I., & Rospocher, M. (2025). On assessing the performance of LLMs for target-level sentiment analysis in financial news headlines. *Algorithms*, *18*(1), 46. https://doi.org/10.3390/a18010046

[39] Liu, D., Zhou, X., & Li, Y. (2025). Balancing performance and cost of LLMs in a multi-agent framework for BIM data retrieval. *Architectural Engineering and Design Management*. Advance online publication. https://doi.org/10.1080/17452007.2025.2456768

[40] Hsu, E., & Roberts, K. (2025). LLM-IE: A Python package for biomedical generative information extraction with large language models. *JAMIA Open*, *8*(2), ooaf012. https://doi.org/10.1093/jamiaopen/ooaf012

[41] Fakhoury, S., Naik, A., Sakkas, G., Chakraborty, S., & Lahiri, S. K. (2024). LLM-based test-driven interactive code generation: User study and empirical evaluation. *IEEE Transactions on Software Engineering*, *50*(9), 2254–2268. https://doi.org/10.1109/TSE.2024.3428972

[42] Jiang, G., Ma, Z., Zhang, L., & Chen, J. (2024). EPlus-LLM: A large language model-based computing platform for automated building energy modeling. *Applied Energy*, *367*, 123431. https://doi.org/10.1016/j.apenergy.2024.123431

[43] Montpetit, A., Phan, A., Khullar, S., Cayrol, R., Quoc-Huy Trinh, V., Knapp, D., & Khellaf, A. (2025). 15 identifying the optimal architecture for using local large language models (LLM) in autopsy pathology reports. *Laboratory Investigation*, *105*(3), 102237. https://doi.org/10.1016/j.labinv.2024.102237

[44] Wang, Q., & Li, H. (2025). On continually tracing origins of LLM-generated text and its application in detecting cheating in student coursework. *Big Data and Cognitive Computing*, *9*(3), 50. https://doi.org/10.3390/bdcc9030050

[45] Rinder, P., Marcille, T., Sinel-Boucher, P., Cals-Maurette, M., Kanoun, D., Levy, C., . . . , & Heudel, P.-E. (2024). Dynamic projection of medication nonpersistence and nonadherence among patients with early breast cancer. *JAMA Network Open*, *7*(5), e2411909. https://doi.org/10.1001/jamanetworkopen.2024.11909

[46] Tang, H., Zhang, C., Jin, M., Yu, Q., Wang, Z., Jin, X., . . . , & Du, M. (2025). Time series forecasting with LLMs: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter*, *26*(2), 109–118. https://doi.org/10.1145/3715073.3715083

[47] Pahune, S., & Akhtar, Z. (2025). Transitioning from MLOps to LLMOps: Navigating the unique challenges of large language models. *Information*, *16*(2), 87. https://doi.org/10.3390/info16020087

[48] de la Torre, C. A., Bradley, B. A., Lee, R. L., Tiwari, A., Wotherspoon, L. M., Ridden, J. N., & Kaiser, A. E. (2024). Analysis of site-response residuals from empirical ground-motion models to account for observed sedimentary basin effects in Wellington, New Zealand. *Earthquake Spectra*, *40*(4), 2475–2503. https://doi.org/10.1177/87552930241270562

[49] Triem, H., & Ding, Y. (2024). "Tipping the balance": Human intervention in large language model multi-agent debate. *Proceedings of the Association for Information Science and Technology*, *61*(1), 361–373. https://doi.org/10.1002/pra2.1034

[50] Tobar, C. G. R., Urmendiz, Y. D. M. M., Vallejo, M. A., Manquillo, D. F., Castaño, V. E. N., Caicedo, A. I. O., . . . , & Cuellar, R. A. D. (2024). Immunomodulatory effect of Tityus sp. in mononuclear cells extracted from the blood of rheumatoid arthritis patients. *Journal of Venomous Animals and Toxins including Tropical Diseases*, *30*, e20230064. https://doi.org/10.1590/1678-9199-JVATITD-2023-0064

[51] Zhou, Y., Zou, J., Zhang, C., Li, D., Ma, L., & Li, M. (2024). Study on grinding removal mechanism and subsurface damage of bionic layered graphene ceramic matrix composites. *The International Journal of Advanced Manufacturing Technology*, *130*(7–8), 3837–3849. https://doi.org/10.1007/s00170-023-12897-7

[52] Wang, J., Zhu, M., Li, Y., Li, H., Yang, L., & Woo, W. L. (2024). Detect2interact: Localizing object key field in visual question answering with LLMs. *IEEE Intelligent Systems*, *39*(3), 35–44. https://doi.org/10.1109/MIS.2024.3384513

[53] Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *SN Computer Science*, *5*(5), 636. https://doi.org/10.1007/s42979-024-02963-6

[54] Peng, Y., Sakai, Y., Funabora, Y., Yokoe, K., Aoyama, T., & Doki, S. (2025). Funabot-sleeve: A wearable device employing McKibben artificial muscles for haptic sensation in the forearm. *IEEE Robotics and Automation Letters*, *10*(2), 1944–1951. https://doi.org/10.1109/LRA.2025.3528229

[55] Mao, Z., Kobayashi, R., Nabae, H., & Suzumori, K. (2024). Multimodal strain sensing system for shape recognition of tensegrity structures by combining traditional regression and deep learning approaches. *IEEE Robotics and Automation Letters*, *9*(11), 10050–10056. https://doi.org/10.1109/LRA.2024.3469811

[56] Mao, Z., Hosoya, N., & Maeda, S. (2024). Flexible electrohydrodynamic fluid-driven valveless water pump via immiscible interface. *Cyborg and Bionic Systems*, *5*, 0091. https://doi.org/10.34133/cbsystems.0091

[57] Lau, S. L. H., Lim, J., Chong, E. K. P., & Wang, X. (2023). Single-pixel image reconstruction based on block compressive sensing and convolutional neural network. *International Journal of Hydromechatronics*, *6*(3), 258–273. https://doi.org/10.1504/IJHM.2023.132303

[58] Verma, H., Siruvuri, S. D. V. S. S. V., & Budarapu, P. R. (2024). A machine learning-based image classification of silicon solar cells. *International Journal of Hydromechatronics*, *7*(1), 49–66. https://doi.org/10.1504/IJHM.2024.135990

[59] Alawi, O. A., Kamar, H. M., Shawkat, M. M., Ani, M. M. A., Mohammed, H. A., Homod, R. Z., & Wahid, M. A. (2024). Artificial intelligence-based viscosity prediction of polyalphaolefin-boron nitride nanofluids. *International Journal of Hydromechatronics*, *7*(2), 89–112. https://doi.org/10.1504/IJHM.2024.138261

[60] Li, J., & Yang, S. X. (2025). Digital twins to embodied artificial intelligence: Review and perspective. *Intelligence & Robotics*, *5*(1), 202–227. https://doi.org/10.20517/ir.2025.11

[61] Yuan, F., Huang, X., Zheng, L., Wang, L., Wang, Y., Yan, X., . . . , & Peng, Y. (2025). The evolution and optimization strategies of a PBFT consensus algorithm for consortium blockchains. *Information*, *16*(4), 268. https://doi.org/10.3390/info16040268

[62] Wang, X., & Reynolds, B. L. (2024). Beyond the books: Exploring factors shaping Chinese English learners' engagement with large language models for vocabulary learning. *Education Sciences*, *14*(5), 496. https://doi.org/10.3390/educsci14050496

[63] Yao, J., & Yuan, B. (2024). Research on the application and optimization strategies of deep learning in large language models. *Journal of Theory and Practice of Engineering Science*, *4*(05), 88–94. https://doi.org/10.53469/jtpes.2024.04(05).12

[64] Ko, J., Kim, S., Chen, T., & Yun, S.-Y. (2024). DistiLLM: Towards streamlined distillation for large language models. *arXiv Preprint: 2402.03898*.

[65] Salemi, A., Kallumadi, S., & Zamani, H. (2024). Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 752–762. https://doi.org/10.1145/3626772.3657783

[66] Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Yu, C. H., . . . , & Sheng, Y. (2024). SGLang: Efficient execution of structured language model programs. In *38th Conference on Neural Information Processing Systems*, 1–27.

[67] Diao, H., Cui, Y., Li, X., Wang, Y., Lu, H., & Wang, X. (2024). Unveiling encoder-free vision-language models. *arXiv Preprint: 2406.11832*.

[68] Gürcan, Ö. (2024). LLM-augmented agent-based modelling for social simulations: Challenges and opportunities. In *Hybrid Human AI Systems for the Social Good: Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence*, 134–144. https://doi.org/10.3233/FAIA240190

[69] Yang, Y., Sun, F.-Y., Weihs, L., Vanderbilt, E., Herrasti, A., Han, W., . . . , & Clark, C. (2024). Holodeck: Language guided generation of 3D embodied AI environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16277–16287. https://doi.org/10.1109/CVPR52733.2024.01536

[70] Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., . . . , & Zhu, S.-C. (2024). The Tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, *34*, 12–22. https://doi.org/10.1016/j.eng.2023.07.006

[71] Dai, N., Zhang, K., Zhang, F., Li, J., Zhong, J., Huang, Y., & Ding, H. (2025). AI-assisted flexible electronics in humanoid

robot heads for natural and authentic facial expressions. *The Innovation*, *6*(2), 100752. https://doi.org/10.1016/j.xinn.2024.100752

[72] Yang, Y., Jia, B., Zhi, P., & Huang, S. (2024). PhyScene: Physically interactable 3D scene synthesis for embodied AI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16262–16272. https://doi.org/10.1109/CVPR52733.2024.01539

How to Cite: Yuan, F., Jiang, H., Guan, H., & Peng, Y. (2025). Frontier Exploration of the Fusion of Embodied Intelligence and Large Language Models. *Smart Wearable Technology*. https://doi.org/10.47852/bonviewSWT52025965