Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29102

Research on the Technological Evolution Path and Realistic Dilemma of Large Language Models

Yuyun Xiang

School of Optoelectronic Information Science and Engineering, Hangzhou Dianzi University, Hangzhou, China
22320322@hdu.edu.cn

Abstract. With the rapid advancement of deep learning, artificial intelligence is shifting from traditional rule- and statistics-based methods to frameworks centered on deep learning. Large-scale pre-trained language models, with their broad applicability, have achieved notable gains in education, healthcare, office work, and other fields. However, at present, these models still face challenges such as illusion generation, lagging knowledge update and ethical risks. This paper aims to discuss the development and application of the large language model based on Transformer architecture, and analyze its performance in various fields and the main problems it faces. Through combing and analyzing the relevant literature, this paper mainly explores the model development, pre-training and fine-tuning strategies based on Transformer architecture, as well as application examples. The results indicate that the large language model has obvious advantages in generating and understanding, but there are still obvious shortcomings in dealing with ethics, safety and timeliness. Future research should emphasize multi-modal integration, value alignment, and efficiency enhancement, ensuring that the technology can be used safely and reliably across various fields, while maximizing social benefits via the effective interplay of technology, ethics, and law.

Keywords: Large Language Models (LLMs), Transformer Architecture, Pre-training and Fine-Tuning, Applications of Large Language Models

1. Introduction

In recent years, the large language model (LLM) has made remarkable progress in the natural language processing, especially in the tasks of text generation, semantic understanding and logical reasoning. However, despite the continuous optimization of model architecture and training strategies, the current large language model still faces problems such as poor reliability, insufficient controllability and false information generation. While the Transformer architecture and pre-training and fine-tuning paradigm improve the model's ability, in practical application, it still exposes its limitations like poor universality and insufficient stability [1,2]. This paper explores the development process, application expansion and challenges of the large language model, discusses its limitations and potential in practical application, and puts forward the future development direction. Through the analysis of Transformer architecture and the basic framework of large

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29102

language model, it addresses key issues in technological evolution, particularly overcoming technical bottlenecks and enhancing reliability and controllability. Besides, the optimization paths of multi-modal integration, value alignment, and efficiency are discussed to enhance the safety, reliability, and inclusiveness of technology. By reviewing literature and analyzing cases, this paper summarizes the technical progress, application effect, and limitations of large language models, provides guidance for their optimization and use, and promotes the field's development in a safer, more reliable, and more inclusive direction.

2. Technical foundations and evolution of large language models

2.1. Key technical architecture

The two pillars of large language model are the Transformer infrastructure and the training paradigm of pre-training and fine-tuning, respectively, and their joint action realizes the model's ability to master complex languages and adapt to various scenarios.

2.1.1. Transformer architecture design

Transformer architecture combines self-attention mechanism and feedforward neural network to solve the computational bottleneck problem faced by traditional models when dealing with long sequences. In contrast to recurrent neural networks (RNNs), Transformers enable parallel processing, significantly enhancing training efficiency. Its architecture comprises an encoder and a decoder, each made up of multiple layers, with self-attention mechanisms and feedforward neural networks integrated across the layers [3]. The self-attention mechanism, the pivotal innovation of the Transformer, enables each word to be represented by its intrinsic information along with its interactions with other words in the input sequence, thus obtaining more comprehensive contextual information. The vector of each input word will interact with the three trained matrix queries, keys and values, thus generating the corresponding Q, K and V vectors. By calculating the dot product of q and k, the model can evaluate the correlation between each word and other words. Then, these correlations are normalized using softmax to obtain the attention weight for each word, which is combined with the V vector to produce the contextual representation of each word. Through this process, it captures the complex dependence between words, thus better understanding the global information and intuitively demonstrating the principle of self-attention mechanism, as shown in Figure 1.

DOI: 10.54254/2755-2721/2026.TJ29102

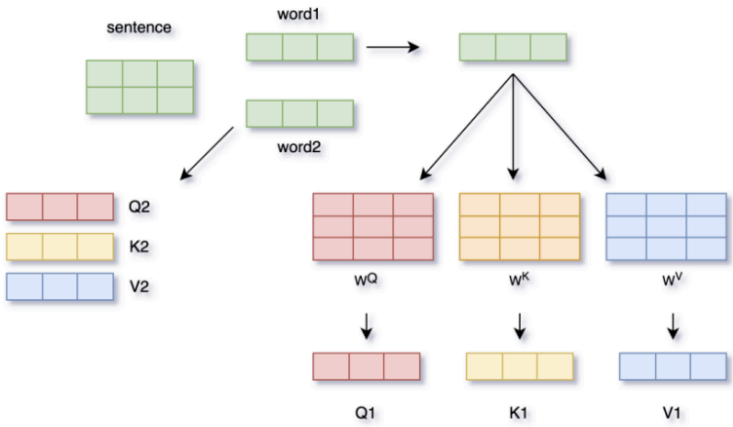

Figure 1. Diagram of QKV vector generation process in self-attention

The further expansion of the self-attention mechanism is the multi-head attention mechanism. By calculating multiple attention heads in parallel, it enables the model to analyze the input information from different angles, and improves the multi-dimensional understanding of the context. Each attention head independently calculates an attention weight, and finally, these results are spliced together and the final output is obtained via linear transformation. Figure 2 shows how multi-head attention integrates the outputs of each head to produce the final result. By boosting expressive capability and accelerating computation, multi-head attention is especially suited for large-scale data processing, establishing its central role in the Transformer architecture.

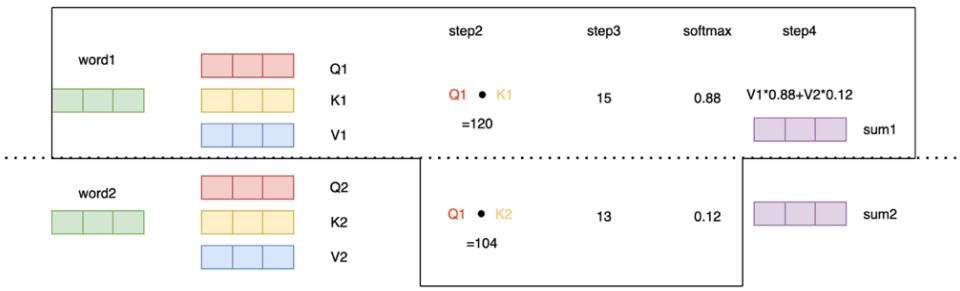

Figure 2. Self-attention weight calculation and context vector generation

In addition, while the decoder resembles the encoder in structure, it also incorporates a linear layer and a Softmax layer. The linear layer combines the output of the encoder with the input of the decoder, and the Softmax layer is used to generate the probability distribution of the next word. Through this mechanism, the decoder generates words sequentially until the sentence is complete. By combining self-attention and multi-head attention, the Transformer architecture markedly enhances large language model performance, improving both expressiveness and adaptability.

2.1.2. Pre-training and fine-tuning paradigm

Pre-training and fine-tuning are the key features of large language model. In the pre-training stage, the model learns language rules through massive unsupervised data. In the fine-tuning stage, the model is adjusted on the specific task data, thus improving the task adaptability.

DOI: 10.54254/2755-2721/2026.TJ29102

Initially, the pre-training technology was mainly adopted in the field of computer vision, especially under the impetus of ResNet and other networks. The pre-training method extracted features from a large number of unlabeled data through self-supervised learning, and achieved remarkable results [4,5]. With its success, the pre-training technology has been introduced into the field of natural language processing (NLP) and has quickly become the core component of the large language model. Traditional pre-training methods emphasizes word vector learning but struggles with polysemy and deep semantic understanding. To solve these problems, ELMo technology proposed in 2018 introduced context-based dynamic word vectors, which greatly improved the model's ability to deal with lexical ambiguity and provided a richer semantic foundation for the subsequent fine-tuning stage [6].

In the fine-tuning stage, the model was trained on data from a specific task to further optimize its performance on that task. Fine-tuning enhances task accuracy while also enabling the model to adapt to the requirements of various domains. As a special supervision and fine-tuning method, instruction fine-tuning has been widely used. It guides the model to comprehend and align with task intentions, thereby enhancing accuracy and robustness on specific instructions. With the introduction of human feedback reinforcement learning (RLHF) algorithm, fine tuning technology has been further enhanced. Through RLHF, the model can adjust its behavior based on human feedback, ensuring alignment with ethical standards and human values, particularly when addressing complex moral issues [7].

2.2. Typical model evolution

At present, large language models mainly follow two development paths: closed-source super-large models and open-source lightweight models. The closed-source super-large models are trained by tech giants using massive data and computing power, offering high-performance black-box API services. However, open source lightweight models have fewer parameters, greater transparency and auditability, and support private deployment and free modification.

2.2.1. Main large language models and their types

The closed-source super-large model and the open-source lightweight model signify two development paths for current large language models. The closed-source super-large model, exemplified by the GPT series, was first introduced by the American company OpenAI. GPT-1, proposed in 2018, was based on the Transformer architecture and consisted of 12 decoders as a unidirectional pre-trained language model. In 2019, by increasing the number of decoders and expanding the training corpus, OpenAI scaled GPT-2 to 1.5 billion parameters, further enhancing its performance. GPT-3, released in 2020, increased the parameter count by two orders of magnitude and further improved tasks such as dialogue generation and machine translation. In 2022, OpenAI added RLHF to GPT-3 and launched InstructGPT

building on GPT-3, OpenAI introduced RLHF and launched InstructGPT, which greatly improved the model's ability to follow instructions [8].

In contrast, open-source lightweight models emphasize technological openness and customization, as well as community-driven development, thus promoting the popularization and commercialization of large language models. The LLaMA series, published by Meta, is a widely recognized example. In 2023, Meta released LLaMA-1, primarily for text generation, followed a few months later by LLaMA-2, which featured an open-source commercial license, incorporated RLHF, and used a dataset of 2 trillion tokens, achieving strong performance [9]. In 2024, Meta

DOI: 10.54254/2755-2721/2026.TJ29102

introduced LLaMA-3, trained on 15 trillion tokens across more than 30 languages, enhancing the model's diversity and adaptability. Thus, closed-source models drive technological advancement with top performance and reliability, while open-source models expand adoption through customization and community innovation. Together, they compete and complement each other, accelerating the development of large language model technology and promoting its practical application.

2.2.2. Multimodal large language models and their extensions

Multimodal large language models greatly enhance understanding and reasoning in complex scenarios by integrating text, image, audio, and other data types. Typical examples include GPT-4V and Gemini. GPT-4V adds image input to GPT-4, enabling it to analyze visual content and combine it with text to generate relevant responses. By contrast, Gemini can processes text, image, audio, and video at once, gaining a more complete understanding across modalities. For instance, generating summaries from audio or making inferences by combining images with text. Multimodal models improve the efficiency of handling diverse information and support complex applications such as autonomous driving, medical image analysis, and content creation. Their ability to integrate multiple modalities reflects the evolution of large language models toward broader, more intelligent applications and drives further technological innovation.

3. Essential competencies and evaluation of large language models

3.1. Key competence and basic application

The key competencies of large language model mainly include text generation, semantic understanding, and logical reasoning. Among these, text generation is the most intuitive capability. By training on hundreds of billions or even trillions of text data, the model can generate coherent and semantically relevant content based on logical relationships. For example, Shenzhen University applied LLMs to develop Cooper's empirical learning model. Using KIMI's large language model, students are assisted in mastering knowledge in cryptography, network security, and system security [10]. On this basis, the semantic understanding enables the model to analyze and summarize large volumes of data, capturing nuanced meaning across complex contexts. This capability is especially evident in tasks like translation and text-based problem prediction. In medical question-answering settings, the model can comprehend patients' non-technical symptom descriptions and relate them to relevant medical concepts, providing preliminary health consultation guidance. Furthermore, logical reasoning enables the model to tackle structured, rule-based problems using known information, extending beyond basic text generation. In numerical reasoning, the Wallace model can handle calculations using the DROP dataset. For logical reasoning, Rajpurkar constructed a logistic regression model using the SQuAD dataset, demonstrating that the model possesses certain reasoning abilities, though its overall performance still lags behind humans [11]. Thus, text generation, semantic understanding, and logical reasoning constitute the core capabilities of large language models in complex scenarios. Through the synergy of these abilities and integration with multimodal inputs, the models support advanced functions such as context learning and tool invocation, enabling the transition from perception to cognitive decision-making.

DOI: 10.54254/2755-2721/2026.TJ29102

3.2. Extended capabilities and practical adaptation

In the process of constantly optimizing the architecture and algorithm, the LLMs gradually develop new capabilities such as context learning, tool calling and multi-round dialogue. In the medical field, in-context learning allows the model to handle new tasks using just a few examples or prompts, without extensive labeled data. This ability also helps the medical model integrate entity recognition results into the knowledge graph by analyzing patients' medical records, test results, and literature data, thereby providing more accurate medical answers and diagnostic or treatment suggestions [12]. Meanwhile, the tool-use capability allows the model to dynamically access external information sources, like retrieving real-time data via web searches, database queries, or professional APIs. This overcomes training data limitations and boosts the model's adaptability in practical applications. For example, when querying the Nobel Prize winners in 2023 or the latest scientific research achievements, the model can retrieve and integrate relevant information in real time and provide accurate answers. In addition, the ability of multi-round dialogue enables the model to maintain coherent and contextually relevant interactions. The model can memorize key information in the conversation, analyze users' preferences and interests, and generate answers that meet users' needs based on the previous content. This ability is especially important in intelligent customer service, educational counseling and personal assistant, which makes the user experience more natural and personalized. Through architectural optimization and algorithmic iteration, building on its core capabilities, the LLM has further developed extended functions such as in-context learning, tool use, and multi-turn dialogue, transforming it from a static knowledge base into an intelligent agent capable of dynamic interaction, autonomous decision-making, and task execution.

3.3. Evaluation methods and benchmark frameworks

LLM training depends on vast amounts of data, which may contain inaccurate information or ethical issues, making comprehensive evaluation essential. Evaluations measure model performance through automated metrics, human assessment, and standard tests. In terms of automatic indicators, the degree of confusion (PPL) is used to measure the fluency and rationality of text generation, which reflects the uncertainty when the model generates text, and its formula is as follows.

$$PP(W)=P(w_1,w_2,...,w_N)^{-1/N}=2^{H(p)} \tag{1}$$

where PP(W) denotes the perplexity of text sequence W; lower values indicate text closer to human language. In chatbot evaluation, low-perplexity sentences are coherent and grammatically correct, thus indicating the model's text generation quality. In chatbot evaluation, low-perplexity sentences exhibit coherence and correct grammar, serving as a metric for the model's text generation performance.

In addition, manual evaluation measures the normality, fluency, relevance, creativity, ethics, and fairness of the model's output via subjective scoring. Specifically, normality checks whether the model solves the problem, fluency assesses the text's flow, relevance evaluates whether the content matches the context, and creativity measures the novelty, ethics, and fairness of the text to ensure it conforms to moral standards. Consequently, manual evaluation can compensate for the limitations of automated metrics in capturing these aspects. Moreover, standardized tests further evaluate model performance via preset benchmark suites [13,14]. For example, GLUE contains nine natural language understanding tasks, while MMLU covers 57 topics, testing the model's accuracy across different knowledge domains using multiple-choice questions. Such tests are widely used to assess

DOI: 10.54254/2755-2721/2026.TJ29102

the capabilities of closed-source large models. However, automated metrics and manual scoring assess performance only on specific tasks and cannot fully capture the model's thinking ability. As standardized testing is generally carried out in static environments, it may overlook complex real-world scenarios and some latent capabilities or risks. As a result, evaluations should be combined with real-world application scenarios to more accurately reflect the model's performance and security risks.

4. Existing challenges and future prospects

4.1. Technical limitations and performance risks

Large language models face risks related to knowledge timeliness, hallucinations, and data security. Knowledge timeliness refers to the limitation that the model can only access information contained in its training data. For example, ChatGPT-3.5's knowledge is limited to events up to September 2021, making it unable to provide information on subsequent events or recent developments. Even online, the model is limited by the availability and quality of data. Hallucinations occur when the model generates false or inaccurate information, such as nonexistent references, which is especially problematic in professional domains. Data security risks arise from sensitive information potentially included in the training data, and users may inadvertently expose private or restricted data through specific prompts. These issues limit the application of large language models in high-reliability scenarios and define their technical boundaries.

4.2. Ethical norms and social challenges

While advancing social development and technological innovation, large language models also pose ethical and social challenges, including algorithmic bias, privacy leakage, and misuse. Algorithmic bias arises from historical biases or unbalanced training data, thus leading to unfair outputs when the model handles sensitive topics such as gender, race, or age. For example, in recruitment or scoring systems, the model may favor certain groups. And privacy leakage occurs when the model inadvertently exposes personal or sensitive information contained in the training data, posing legal and ethical risks. Misuse happens when the model spreads false or misleading information or aids cheating and fraud, worsening its negative impact. To address these challenges, a robust technical ethics framework should be applied, including data anonymization, bias detection, and legal and social oversight, ensuring the safe and compliant deployment of LLMs in practical applications.

4.3. Development trend and future direction

At present, large language models an communicate using natural language. Although the output is occasionally inaccurate, its ability has exceeded the initial expectations. The future development will focus on several key areas, including human alignment, multimodal processing, embodied intelligence and social application optimization. In terms of human alignment, the model will ensure that the output conforms to human values and actual needs through external feedback. Multimodal processing requires the model not only to understand text but to process images, audio, video, and other data. Embodied intelligence emphasizes the interaction between the model and the external environment, so that it can reflect and correct its mistakes independently and enhance its practical application capabilities. Social application optimization focuses on the coordination of technology, ethics and law to ensure the safe and effective application of the model in education, medical care,

DOI: 10.54254/2755-2721/2026.TJ29102

finance and other fields. Future research will focus on these directions to promote the further development of large language models in intelligence and social adaptability.

5. Conclusion

This study indicates that large language model technology has developed from early self-supervised learning to the pre-training and fine-tuning paradigm through continuous updates and optimization. Leveraging the advantages of the Transformer architecture, model performance has gradually improved. With the expansion of datasets and functional enhancements, these models have been widely applied in daily life, healthcare, and education, markedly transforming people's work and lifestyle. In the future, effectively regulating the use of large language models and ensuring their safety and compliance across multiple scenarios will be a key focus. The further development of large language models is expected to play an increasingly critical role in social progress and human well-being.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
[2] Li, Z., Fan, Y., & Wu, X. (2020). Review of pre-training technology for natural language processing. Computer Science, 47(03), 162–173.
[3] Bi, X. (2025). Integrating qualitative and quantitative analysis with "computational understanding" — Research on evaluating teaching big data based on Transformer architecture. Intelligent Society Research, 2025(03).
[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
[5] Zhang, Q., Wang, Y., Wang, H., et al. (2024). Summary of fine-tuning technology of large language model. Computer Engineering and Application, 60(17), 17–33.
[6] Rajani, N. F., & Mooney, R. (2018). Stacking with auxiliary features for visual question answering. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2217–2226.
[7] Xu, Y., Wang, M., Chenglong, et al. (2025). Pipe-RLHF: Parallel acceleration framework for computing pattern awareness. Computer Research and Development, 62(06), 1513–1529.
[8] Sun, Q., Shi, J., Pei, L., et al. (2024). Fine-tuning optimization method of LLaMA2 model based on transfer learning. Software Engineering, 27(09), 32–36.
[9] Lu, Y., Yu, J., Chen, P., et al. (2023). Educational application and prospect of generative artificial intelligence—Taking ChatGPT system as an example. China Distance Education, 43(04), 24–31+51.
[10] Yang, B., Sun, X., Dang, J., et al. (2023). Named entity recognition method of large language model for medical question answering system. Computer Science and Exploration, 17(10), 2389–2402.
[11] Ni, R., Xiao, D., & Gao, P. (2024). Research on the evaluation of basic logical ability of large-scale pre-training language model. Journal of Qufu Normal University (Natural Science Edition), 50(03), 89–95.
[12] Yan, Q., Zhuang, J., & Zhong, S. (2025). Application of Cooper's empirical learning model based on large language model in the course of introduction to computer security. Computer Education, 2025(8), 59–63.
[13] Wang, A., Singh, A., Michael, J., et al. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355.
[14] Hendrycks, D., Burns, C., Basart, S., et al. (2020). Measuring massive multitask language understanding. In Proceedings of International Conference on Learning Representations.