Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.25586

Large Language Models and Their Evolution

Zihan Zhou

School of Computer Science and Engineering, Southeast University, Nanjing, China
zh.zhou.seu@gmail.com

Abstract. With the rapid advancement of artificial intelligence (AI), large language models (LLMs) have become the foundational infrastructure for natural language processing (NLP) research and industrial applications. By leveraging massive parameters and vast pre-training data, LLMs have significantly enhanced text understanding, generation, and cross-modal reasoning capabilities. This paper systematically reviews the technical evolution of LLMs from n-gram statistical models to the Transformer architecture, based on five key review papers. It analyzes training and alignment paradigms such as "pre-training & fine-tuning" and "RLHF/RLAIF", as well as the exponential parameter expansion driven by the Scaling Law. Furthermore, we summarize the latest application progress of LLMs in code generation, intelligent customer service, medical and legal assistance, and other fields, and analyze the challenges they face in terms of data privacy, model bias, hallucination phenomena, and energy consumption. Finally, this paper proposes four research priorities for the future: first, leveraging explainable mechanisms to enhance model transparency; second, strengthening value alignment and security controls; third, exploring green and efficient model compression and inference schemes; and fourth, leveraging interdisciplinary collaboration to build the next generation of general-purpose intelligent systems that are both fair and sustainable.

Keywords: Large Language Models, Transformer, Scaling Law, Pre-training & Fine-tuning, Prompt Engineering

1. Introduction

In recent years, the exponential growth of artificial intelligence (AI) capabilities has been most evident in breakthroughs in natural language processing. At the heart of these breakthroughs are large language models (LLMs) such as BERT and GPT series, which have greatly improved human-computer interaction and content generation standards. Leveraging massive computing resources and vast amounts of text data, these models demonstrate unprecedented capabilities, including understanding context, generating coherent language, and performing tasks in diverse domains [1].

While celebrating these technological milestones, it is equally important to gain a deeper understanding of the evolution and underlying principles of LLMs. From early statistical language modeling to contemporary transformer-based architectures, the field has undergone a radical transformation, fundamentally changing the way models process language data [2].

© 2026 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: 10.54254/2755-2721/2025.25586

LLM research has not only brought theoretical innovations in academia, but also had a wide-ranging impact in industry, such as intelligent customer service, search engine optimization, code generation, and medical diagnosis assistance. At the same time, it has also sparked widespread discussions on algorithm transparency, data privacy, model bias, and sustainability.

This paper provides a systematic review of the development of LLMs, focusing on major evolutions in model structure, training paradigms, and application extensions, while also exploring the core challenges and future research directions. The paper is divided into three main parts: the first part traces the technological evolution; the second part discusses the evolution of training and alignment strategies; and the third part analyzes the application scenarios of LLMs and the social challenges they bring, concluding with a look ahead to future research.

2. Summary

2.1. Evolution and principles of LLMs

The objective of language modeling is to predict the probability of the next word in a sentence based on the context. Early language modeling primarily used n-gram statistical models, which are based on the Markov assumption that a word depends only on a fixed number of preceding words. Although simple and efficient, n-gram models suffer from severe data sparsity issues and context window limitations, making it difficult to capture long-range dependencies. To address these issues, neural network methods were introduced, starting with feedforward neural network language models (NNLMs), which evolved into recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). LSTMs effectively mitigated the vanishing gradient problem by introducing forget, input, and output gates, enabling the retention of longer contextual information. However, due to the sequential structure of RNNs, their training efficiency remained low [2].

A milestone development in the field of natural language processing occurred in 2017 when the Transformer architecture was proposed, which revolutionized the field of language modeling. Its core technology, the self-attention mechanism, enables the model to calculate the relationships between all words in a sequence at each layer, eliminating the need for recursive structures and achieving efficient parallel processing and significantly improved long-term dependency handling capabilities. In addition to the architectural leap brought by Transformers, another critical advancement lies in the introduction of positional encoding, which enables the model to capture sequence order without relying on recurrence. This is crucial because, unlike RNNs, Transformers process tokens in parallel and do not have an inherent notion of token order. The sinusoidal positional encoding scheme allows the model to incorporate relative and absolute positions into attention calculations, thus preserving syntactic structure and temporal dependencies across long texts [3].

Moreover, recent innovations have explored alternatives to standard self-attention to improve efficiency and scalability. Notably, sparse attention mechanisms, such as those implemented in models like Longformer and so on, enable the handling of longer sequences with reduced computational overhead by restricting attention to local and global patterns selectively [4]. These methods mark an important direction in improving the usability of LLMs in scenarios with long-form documents or multi-document inputs. This innovation forms the foundation of almost all contemporary LLMs, including GPT-3 and Google's BERT, enabling models to be trained at unprecedented scales of data and parameters, ushering in a new era of large-scale pre-trained models [5].

DOI: 10.54254/2755-2721/2025.25586

From GPT-2's 1.5 billion parameters to GPT-3's 175 billion, and then to PaLM and GPT-4's parameters in the trillions, the development of LLMs exhibits a typical "scaling law" phenomenon. Research by OpenAI has found that increasing model size, data volume, and training computing resources can lead to continuous performance improvements, but the marginal benefits gradually decrease [6]. This also indicates that as long as training resources are sufficient, continuing to increase model size can yield sustained gains, albeit with diminishing marginal benefits. This principle has driven the rapid expansion of model scale, with GPT-4 already containing trillions of parameters, marking the arrival of the era of ultra-large-scale models [1].

2.2. Training paradigms and alignment strategies

The current mainstream training paradigm adopts a two-stage approach of "pre-training & fine-tuning". In the pre-training stage, large-scale unlabeled data is used for self-supervised training. The GPT series uses causal language modeling, which predicts the next word based on the preceding text; BERT uses masked language modeling, which randomly masks words in a sentence and trains the model to fill in the blanks. These tasks help models learn general language knowledge. Subsequently, in the fine-tuning stage, these models undergo supervised training with a small amount of labeled data to adapt to specific downstream tasks such as question answering, classification, and summarization, significantly improving their practical application capabilities and efficiency [7]. This training paradigm enables a unified model to generalize to multiple tasks and achieve good results in low-resource (few-shot) and zero-resource (zero-shot) scenarios. This paradigm significantly reduces the annotation cost of various NLP tasks and demonstrates strong generalization capabilities in few-sample and even zero-sample settings.

In traditional training methods, models optimize language modeling objectives and do not necessarily generate answers that are valuable to users. Therefore, alignment strategies have been proposed to ensure that the output of LLMs is consistent with human values and preferences. Reinforcement learning with human feedback (RLHF) has played a key role in optimizing models such as ChatGPT, guiding their outputs to be safer, more reliable, and contextually appropriate. The RLHF training process consists of three stages: supervised fine-tuning, reward model training, and reinforcement learning. Specifically, human evaluators first rank the outputs, then use these rankings to construct a reward model, and finally adjust the language model through reinforcement learning. Given the rising cost of labeling, automated methods such as reinforcement learning with AI feedback (RLAIF) are also gaining attention. This method uses another LLM to score answers and train the reward model, automating the training loop and improving alignment efficiency [7].

Another key frontier in training strategies is the emergence of instruction-tuned models. These models are fine-tuned using large-scale instruction datasets—collections of prompts and corresponding desired outputs—to enhance their ability to follow natural language commands [8]. Notable examples include InstructGPT and FLAN-T5, which significantly improve user satisfaction by generating outputs that are more aligned with human intentions and better formatted for end-user consumption. Beyond instruction tuning, continual learning and model updating have gained attention. Rather than retraining models from scratch, incremental updates with new data allow LLMs to stay current with evolving knowledge, a necessity in domains such as finance or medicine where information changes rapidly [9]. This shift toward "lifelong learning" in LLMs also poses challenges in maintaining alignment and avoiding catastrophic forgetting, which is now an active area of research.

Lastly, alignment is increasingly intertwined with interpretability. Recent efforts aim to incorporate attribution methods (e.g., attention visualization, integrated gradients) during training to

DOI: 10.54254/2755-2721/2025.25586

encourage transparent reasoning paths, making it easier to audit model decisions and identify sources of misalignment or bias [10].

2.3. Diversified applications and social impact

LLMs are disrupting multiple industries. In software engineering, models like Codex and Copilot enhance productivity by providing code completion, debugging assistance, and even automatic defect detection [11]. Additionally, chatbots like ChatGPT redefine customer service by generating human-like conversational interactions, significantly impacting user experience and operational efficiency across industries. Beyond these areas, LLMs are driving technological advancements in education, healthcare, and legal assistance, offering intelligent tools capable of summarizing literature, diagnosing medical conditions from text data, and interpreting legal documents. Furthermore, integrating multimodal capabilities into LLMs has greatly expanded their application scope, enabling unified understanding and generation of text, audio, and visual inputs [1].

In the scientific domain, LLMs have begun to accelerate discovery workflows. For instance, models like Galactica and so on are trained specifically on scientific literature, aiding in knowledge synthesis, hypothesis generation, and even experimental planning [12]. This has empowered researchers across fields like chemistry, biology, and materials science to extract insights from vast corpora of research papers efficiently. Moreover, governments and policy institutions are exploring the use of LLMs for public service enhancements. Applications include automated policy analysis, intelligent document drafting, and real-time translation of public information into multiple languages, which can significantly improve accessibility and civic engagement [13].

Despite the significant benefits of LLMs, their large-scale deployment faces major ethical, privacy, and sustainability challenges. For example, large-scale training data may contain sensitive information, posing a risk of leakage, and researchers are attempting to introduce differential privacy mechanisms to protect data security; historical biases in training data can be inherited or even amplified by models, leading to risks such as gender, racial, cultural discrimination, necessitating the establishment of fairness mechanisms and diversity assessment methods; LLMs exhibit uncontrollable phenomena in content generation, including hallucinations, logical flaws, and inappropriate speech, with controllable generation, refusal mechanisms, and instruction alignment remaining key research priorities; and the expansion of model scale leads to exponential computational demands, posing challenges for green AI, with dedicated AI chips and sparsification techniques identified as potential breakthroughs. In summary, these issues of data bias, model hallucinations, and large-scale energy consumption require targeted research and solutions to ensure that artificial intelligence development remains fair, safe, and environmentally responsible [5].

3. Conclusion

This paper provides a systematic review of the technological evolution and methodological innovations in LLMs, charting their development from early statistical approaches to contemporary large-scale Transformer architectures. We highlighted key breakthroughs in model design, including self-attention mechanisms and scale laws, and traced the evolution of training strategies, progressing from basic pre-training and fine-tuning to sophisticated alignment techniques like RLHF and instruction tuning. Significantly, the paper documented the profound social impact of LLMs, demonstrating their transformative applications across diverse fields such as software engineering, healthcare, education, and legal services, fundamentally altering workflows and capabilities.

DOI: 10.54254/2755-2721/2025.25586

Looking ahead, the responsible advancement of LLMs necessitates concerted efforts across several critical dimensions. Firstly, enhancing model explainability and controllability is paramount for user trust and developer oversight. Secondly, mitigating inherent social biases and ethical risks, including content misuse, demands robust research in bias detection, alignment techniques, and ethical frameworks. Thirdly, addressing sustainability concerns requires innovations in efficient computation, such as model compression, sparsity techniques, energy-efficient hardware, and optimized training paradigms. Finally, fostering interdisciplinary collaboration is crucial, engaging ethicists, sociologists, legal experts, and policymakers alongside technologists to develop global governance standards, privacy safeguards, and responsible deployment guidelines.

Ultimately, the ascent of LLMs signifies not merely a technological leap, but a fundamental shift in societal information processing, cognition, and decision-making. Navigating this transformation requires a comprehensive, ethically grounded, and socially inclusive approach to realize the potential of LLMs while ensuring they become a force for equitable progress and a trustworthy foundation for the intelligent era.

References

[1] Jana S, Biswas R, Pal K, et al. The evolution and impact of large language model systems: A comprehensive analysis [J].Alochana Journal 2024.
[2] Wang Z, Chu Z, Doan T V, et al. History, development, and principles of large language models: an introductory survey [J]. AI and Ethics, 2024: 1-17.
[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
[4] Beltagy I, Peters M E, Cohan A. Longformer: The long-document transformer [J]. arXiv preprint arXiv: 2004.05150, 2020.
[5] Naik D, Naik I, Naik N. Large data begets large data: studying large language models (LLMs) and its history, types, working, benefits and limitations [C] The International Conference on Computing, Communication, Cybersecurity & AI. Cham: Springer Nature Switzerland, 2024: 293-314.
[6] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models [J]. arXiv preprint arXiv: 2001.08361, 2020.
[7] Liu Y, Han T, Ma S, et al. Summary of chatgpt-related research and perspective towards the future of large language models [J]. Meta-radiology, 2023, 1(2): 100017.
[8] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners [J]. arXiv preprint arXiv: 2109.01652, 2021.
[9] Jang J, Ye S, Yang S, et al. Towards continual knowledge learning of language models [J]. arXiv preprint arXiv: 2110.03215, 2021.
[10] Jacovi A, Goldberg Y. Aligning faithful interpretations with their social attribution [J]. Transactions of the Association for Computational Linguistics, 2021, 9: 294-310.
[11] Zheng Z, Ning K, Wang Y, et al. A survey of large language models for code: Evolution, benchmarking, and future trends [J]. arXiv preprint arXiv: 2311.10372, 2023.
[12] Taylor R, Kardas M, Cucurull G, et al. Galactica: A large language model for science [J]. arXiv preprint arXiv: 2211.09085, 2022.
[13] Aoki G. Large Language Models in Politics and Democracy: A Comprehensive Survey [J]. arXiv preprint arXiv: 2412.04498, 2024.