Framework Design and Performance Comparative Analysis of Large Language Models

Mingxuan Deng

School of Engineering Science, Lappeenranta-Lahti University of Technology, Yliopistonkatu 34, 53850 Lappeenranta, Finland
Mingxuan.deng@student.lut.fi

Abstract. With the emergence of Transformer architecture, large language models (LLMs) have made breakthroughs in the fields of language understanding, reasoning, code generation and multimodal interaction. This research systematically sorts out the technical evolution path of mainstream LLM in the past five years, from pre-training paradigm, architecture characteristics, instruction fine-tuning to multi-modal expansion, and analyzes the differences between different models in data sources, training strategies, structural optimization and task performance. On the basis of reviewing a large number of literatures, the main differences between the closed-source model and the open source model in terms of intelligence level, scalability and application potential are summarized, and it is pointed out that problems such as resource efficiency, illusion control, security alignment and cross-modal consistency still constitute the core challenges of current technological development. Through the discussion of the typical benchmark data set and the evaluation index system, this study further reveals the trade-off relationship between model performance, bias, security and interpretability. Finally, this article proposes that in the future, LLM will deepen its development in the direction of efficient architecture, deploy ability, multimodal integration and value alignment, providing reference for subsequent academic research and engineering applications.

Keywords: Large language models; Transformer; Model comparison; Multimodal; Resource efficiency

1 Introduction

Since the introduction of the Transformer model structure in 2018, the development speed of Large Language Models (LLMs) showed an unprecedented explosive growth. Based on large-scale corpus and high-performance computing resources, pre training enables the model to have the ability to understand and generate complex languages. In recent years, with the continuous evolution of parameter size, training strategies, and alignment mechanisms, the performance of LLM has expanded from language generation to a wider range of intelligent tasks such as inference, code writing, knowledge retrieval, and multimodal interaction. The model families represented by ChatGPT, Claude, Gemini, Llama, Mistral, Falcon, DeepSeek, etc. continue to emerge,

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_89

with different technical routes, capability boundaries, and applicable scenarios, forming a diverse intelligent ecosystem. How to systematically compare these models, understand their performance differences, design ideas, and future trends has become an important direction in current artificial intelligence research.

The comparative study of big language models is of great significance. On the one hand, it helps to reveal the differences between different models in terms of data size, parameter quantity, training methods, optimization mechanisms, and secure alignment, thereby helping the academic community understand the formation mechanism of language intelligence. On the other hand, comparative research provides decision-making basis for engineering practice, enabling developers to select the most suitable model for application in fields such as education, scientific research, industrial design, software development, etc. based on task characteristics and resource constraints. In addition, in the context of increasing attention to the ethics and governance of artificial intelligence, in-depth analysis of the performance of various models in terms of security, bias control, information credibility, and multilingual fairness also has a positive impact on responsible AI development.

At present, scholars and enterprises from all over the world have achieved rich results in the research of performance and architecture of large language models. Achiam et al. comprehensively introduced the training methods, architecture improvements, and multimodal input capabilities of GPT-4 in the GPT-4 Technical Report, pointing out that its performance in language understanding, mathematical reasoning, programming, and visual tasks is close to human level, and emphasizing the importance of safe alignment and model control [1]. Anil et al. proposed a model system based on high-quality multilingual data and efficient training strategies in the PaLM 2 Technical Report, which significantly improved the model's generalization ability in logical reasoning and cross language tasks, laying the foundation for the subsequent Gemini series [2]. Meta AI proposed a solution for training high-performance models using publicly available data in LLaMA: Open and Efficient Foundation Language Models, creating an open and shared model ecosystem that enables researchers to replicate top-level performance at a lower cost [3]. The "Mistral 7B" model published by Jiang et al. shows that through architecture optimization and training strategy improvement, small-scale models can surpass large-scale models in multiple benchmark tests, showing the great potential of efficient lightweight models [4]. The DeepSeek AI team proposed a multi-head potential variable attention mechanism (MLA) and a sparse expert mechanism in DeepSeek V2, which achieves a balance of performance and cost by compressing KV cache and dynamically activating parameters, and provides competitive open source results for Chinese teams in the research [5].

This article aims to systematically analyze the differences in architecture differences, training strategies and performance of representative large-scale language models in the past five years and explore the different characteristics of closed-source and open-source methods in terms of intelligence level, scalability and application potential. The following chapters will discuss in detail the principles of model design, technical evolution path, capability comparison and development trend, aiming to reveal the commonalties and differences of contemporary mainstream language models through

literature analysis, and provide systematic reference and inspiration for academic research and engineering applications.

2 Overview and Development History of Mainstream Technologies

2.1 The basic concepts of language models and the formation of pre-training paradigms

The development of large-scale language models stems from the continuous evolution of language modeling methods. Minaee et al. pointed out that from the early n-gram statistical model to the neurolinguistic model, to the pre-training model based on Transformer, the core change is that "the model gradually transitions from local probability estimation to global context modeling" [6]. Transformer's self-attention mechanism significantly improves the long sequence modeling ability through parallelization, so that early pre-training models such as BERT and GPT can obtain common language characterization on large-scale corpus. Han et al. concluded that this paradigm was first carried out unsupervised pre-training on large-scale corpus, and then calibrate the model capability through supervision fine-tuning and human feedback [7]. This process laid an unified training framework for modern LLM. Overall, the advantage of this stage is to achieve cross-task transfer, while the disadvantage is the rapid expansion of the parameter scale and the high training cost.

2.2 Mainstream architecture system: Transformer, instruction fine-tuning and multi-modal expansion

With the expansion of the model, Transformer has gradually become the core architecture of modern LLM. Han et al. pointed out that the self-attention mechanism constitutes the basis of LLM, and the decoder-only architecture has become mainstream due to the generation stability [7]. On the basis of this structure, Instruction Tuning and RLHF further enhance the ability of the model to follow human intentions, evolving from a language model to a system with complex task reasoning capabilities. Multimodal expansion has become a highlight of development in recent years. Yin et al. pointed out in the multimodal review that multimodal LLMs take the language model as the core of reasoning and realize unified sequence input through visual encoder-projection layer-LLM, thus having the ability of mathematical reasoning and cross-modal understanding without OCR [8]. However, Minaee et al. also pointed out that the secondary attention complexity of Transformer leads to a rapid increase in resource consumption of the model in the long sequence and reasoning stages [6], showing the limitations of the structure itself.

2.3 Technological evolution driven by resource efficiency: efficient structure and system optimization

With the further expansion of the model, resource efficiency has become a key factor affecting the trend of LLM technology. Bai et al. pointed out in the system review of resource-efficient LLM that computing, memory, energy consumption and communication delay are becoming the main bottlenecks in the LLM life cycle, and the availability of the model increasingly depends on efficient structural design [9]. Therefore, methods such as sparse attention, MoE expert, KV cache compression, quantification and distillation are widely adopted to reduce resource overhead under the premise of maintaining performance. At the same time, Yin and others pointed out that the multimodal model has a higher overall resource demand due to its more complex structure, which further promotes the necessity of high-efficiency technology [8]. In terms of the overall trend, Minaee et al. believe that LLM is developing from scale-driven to equally important structural innovation and resource optimization [6], forming three parallel routes of capacity expansion, efficient architecture and multimodal intelligence.

3 Model Architecture Comparison and Performance Analysis

3.1 Source, scale and characteristics of the data set

At present, the evaluation of large language models mainly depends on three types of data sets: first, code generation and program reasoning classes, which are used to test the ability of the model to generate code and make logical reasoning; second, natural language understanding classes, covering tasks such as reading comprehension, common sense inference and multiple-choice reasoning; third, open-domain question-and-answer classes, which are used to evaluate the model's mastery of factual knowledge and the accuracy of its answers. These data sets cover different tasks from classification, reading comprehension to code synthesis, and generally give sample scale, task composition and their input and output forms. For example, TruthfulQA contains 817 questions from 38 categories to test whether the model will restate common misunderstandings. The problem design comes from the common human misconception construction strategy [6]; OpenBookQA has about 6,000 questions, and it contains a collection of core facts and supplementary facts and emphasizes the ability to supplement knowledge [6]. Zhao et al. further integrated these benchmarks into comprehensive frameworks such as MMLU, BIG-bench, HELM, etc. to more systematically cover the ability dimensions of reasoning, multilingualism, knowledge, generation, etc. [10]. In terms of data preprocessing, multi-hop reasoning data sets such as HotpotQA provide "gold paragraphs" and key sentence annotations to ensure that the model must answer the question through cross-sentence reasoning rather than surface matching [6]. The source, scale and task characteristics of such data sets together form the basis for subsequent model comparison.

3.2 Evaluation Metrics

The mainstream metrics system consists of Accuracy, F1-score, Exact Match, ROUGE, BLEU, Pass@k, MC1/MC2, etc. Different tasks use different indicators. For example, reading comprehension tasks generally use EM and F1 to measure the accuracy and recall of answers; summary tasks use ROUGE; code generation uses the Pass@k measurement model to generate the proportion of test cases [6]. TruthfulQA uses MC1, MC2 and BLEURT to measure the authenticity of the answer and the quality of information [6]. The HELM framework proposed by Liang et al. extends the traditional indicators, integrating seven types of indicators of accuracy, calibration, robustness, fairness, bias, toxicity and efficiency into the unified assessment system, and measuring 87.5% of the index combinations in 16 core scenarios, so that the model is no longer just accurate. [11]. Zhao et al. pointed out that MMLU and BIG-bench adopt unified Accuracy as the standard for cross-field comparison, which makes the comprehensive capability differences of large models easier to present [10]. The combination of these indicators makes the comparison of LLM no longer limited to single task performance, but presents a multi-dimensional capability structure.

3.3 Analysis of experimental results

Based on the above data sets and indicators, different literatures give a number of representative experimental trends. Zhao et al. reported that GPT-4 achieved an accuracy rate of 86.4% in the MMLU 5-shot setting, significantly exceeding the previous model, indicating the advantage of the large-scale language model in comprehensive subject knowledge [10]. The same paper also pointed out that in BIG-bench, large models exceed the average human level in 65% of tasks, while small models perform almost random in difficult reasoning tasks such as BIG-bench hard, and chain thinking prompts are needed to improve their performance [10]. On the other hand, the results summarized by Minaee et al. show that the emergence of efficient architecture is changing the traditional understanding of "the larger the scale, the better the performance". Mistral-7B surpasses LLaMA-2-13B in multiple benchmarks, and outpers 34B scale model in code, mathematics and reasoning tasks [6]; Guanaco, which uses QLoRA fine-tuning, reaches Chat on the Vicuna benchmark 99.3% of GPT shows that researchers with limited training resources can also acquire the ability to communicate close to the closed source system [6]. PaLM-540B uses 780 billion tokens to train and achieve the most advanced few-shot results on hundreds of tasks, while Flan-PaLM achieves an average performance improvement of about 9.4% through instruction fine-tuning on 473 data sets [6]. In a more systematic comparison, HELM conducts a unified evaluation of 30 models in 42 scenarios, increasing the common evaluation coverage of the model from the previous 17.9% to 96%, and revealing that there is a significant trade-off between performance, bias, toxicity and efficiency. Some high-performance models are insufficient performance in fairness or toxicity indicators [11]. These experimental results together show that large models maintain a lead in comprehensive ability, but through architecture optimization and efficient fine-tuning methods, small and medium-sized models can also achieve high competitiveness in

several tasks. The differences between models are no longer simply determined by scale but show the common role of "scale-architecture-data".

4 Discussion

4.1 Main challenges

The main challenges of the current big language model can be summarized into four aspects. First of all, the computing cost of model training continues to rise. Minaee and others pointed out that mainstream models require tens of billions to trillions of tokens and huge computing power, and the quadratic complexity of traditional Transformer makes long-sequence training extremely expensive [6]. Secondly, the authenticity and security of the model have not been fundamentally improved. The evaluation results of HELM show that the high-performance model still has obvious shortcomings in the dimensions of bias, toxicity and fairness [11]. Thirdly, multimodal models still have structural difficulties in semantic alignment and cross-modal reasoning. Yin et al. pointed out that multimodal LLM is still prone to "modal illusions" and the cross-modal semantic consistency is unstable [8]. Finally, the threshold of model deployment is still high. The storage, bandwidth and reasoning delay problems caused by high parameter scale make it difficult for many models to be deployed on a large scale in edge computing or mobile environments, and there are still contradictions between resource efficiency and practical applications [12]. These problems together constitute the core bottleneck in the context of the continuous growth of the model scale.

4.2 Outlook

Develop towards high efficiency and deployable. Sun et al. pointed out that efficient architecture can significantly reduce costs while maintaining competitive performance, which will become an important trend in future model design [12]. This conclusion is consistent with many resource-efficient LLM reviews, indicating that future research will shift from "expanding scale" to "improving cost-effectiveness and deployability".

Improve authenticity, safety and reliability. Minaee et al. emphasized that hallucinations and lack of fact are still key issues that limit the large-scale application of models, and the reliability of the model needs to be improved through retrieval enhancement, external knowledge integration and verifiable reasoning methods [6]. At the same time, HELM indicates that value alignment and social indicators will be given more attention in the future assessment system [11]. In addition, Yin et al. pointed out that more consistent cross-modal representation and alignment mechanisms are needed in the future to support the development of general intelligence [8].

Move from monomer intelligence to composite system intelligence. It is mentioned in many papers that the future system may be a combination of multiple professional models, retrieval engines, tool modules and planners, so that LLM can change from "a single model to all tasks" to "modular + combination intelligence" to achieve higher interpretability and robustness [6, 12].

5 Conclusion

Through the systematic analysis of large-scale language models in terms of architecture design, training paradigm, performance and resource efficiency in recent years, this article comprehensively sorts out the development vein and research trend of mainstream LLM technology. The development of large-scale language models has shown a trend from "scale-led" to "structural innovation, efficiency optimization and multi-modal integration". On the one hand, the performance of models such as GPT series, PaLM series and Mistral shows that the expansion of the model scale can still bring obvious advantages in general tasks; on the other hand, the open source system and lightweight model can be used in some tasks through sparse experts, linear attention, quantification and distillation technology. The performance of large models is equal to or even surpassed, which shows that the ability of the model is no longer entirely determined by the scale. At the same time, the emergence of HELM and other comprehensive evaluation frameworks emphasizes factors such as bias, fairness, toxicity and efficiency beyond performance, making the comparison of LLM more realistic. However, with the continuous enhancement of the model's ability, its shortcomings in authenticity, reasoning consistency, bias security, multimodal semantic alignment, etc. have also become more and more prominent. As analyzed in Chapter 4, illusions, missing alignment and high resource expenses have become key constraints to further expand the use of the model. These challenges show that the development of LLM has entered a new stage that requires performance, efficiency, security and controllability.

Looking to the future, the evolution of LLM will continue to advance in two directions. First of all, in terms of high efficiency and deployability, sparse architecture, low-complexity attention mechanism, cache optimization, model compression and other technologies will be further mature, making "faster, lighter and more economical" an important standard for model design, providing a wider application space for resource-limited scenarios. Secondly, in terms of reliability and multimodal intelligence, with the development of retrieval enhancement, verifiable reasoning, value alignment, multimodal representation learning and cross-modal consistency technology, LLM is expected to achieve breakthroughs in factuality, security and complex scenario understanding, making it more suitable for real, complex and high-risk environments.

References

1. Achiam, J., et al.: GPT-4 Technical Report. arXiv:2303.08774 (2023)

2. Anil, R., et al.: PaLM 2 Technical Report. arXiv:2305.10403 (2023)
3. Meta AI.: LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 (2023)
4. Jiang, A.Q., et al.: Mistral 7B. arXiv:2310.06825 (2023)
5. DeepSeek-AI.: DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434 (2024)
6. Minaee, S., et al.: Large Language Models: A Survey. arXiv:2402.06196 (2024)
7. Han, S., et al.: A Review of Large Language Models. Electronics **13**(1), (2024)
8. Yin, S., et al.: A Survey on Multimodal Large Language Models. National Science Review (2024)
9. Bai, G., et al.: Beyond Efficiency: A Systematic Survey of Resource-Efficient LLMs. arXiv:2401.00625 (2024)
10. Zhao, W.X., et al.: A Survey of Large Language Models. arXiv:2303.18223 (2025)
11. Liang, P., et al.: Holistic Evaluation of Language Models (HELM). Transactions on Machine Learning Research (TMLR) (2023)
12. Sun, W., et al.: Speed Always Wins: Efficient Architectures for Large Language Models. arXiv:2508.09834v1 (2025)