ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 13 (2026) No. 2, pp. 43-61
DOI:10.33168/JLISS.2026.0203

The Development History, Technological Optimization, and Multi-Domain Applications of Large Language Models

ZhiHao Liu [1,2], WaiYie Leong [2]*

[1]Nantong Institute of Technology, 226000 Nantong, Jiang su, China
[2]INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia
i24028830@student.newinti.edu.my; waiyie.leong@newinti.edu.my (Corresponding author)

Abstract. Large language model is one of the most popular research directions in the world. In view of its rapid development and far-reaching impact on society, this paper systematically reviews the development process, technology optimization and its application of large language model (LLMS). From the early natural language processing (NLP) rule-driven stage to the Transformer architecture revolution based on deep learning, LLMs have achieved a paradigm breakthrough in language understanding through parameter scale expansion and pre-training technology innovation. Large models represented by GPT series and BERT show strong generalization ability and emergent characteristics, but they are faced with challenges such as computational resource consumption, insufficient interpretability and ethical alignment. The optimization technique covers pre-training, cue learning, and fine-tuning, which significantly improves the efficiency of the model in task adaptation. In terms of applications, LLMs have penetrated into education, finance, health care and manufacturing to promote intelligent upgrading, while addressing issues such as data privacy, professional field adaptation and reliability in high-risk scenarios. The rapid development of China's big models (such as Wenxin, Tongyi, DeepSeek, etc.) marks the rise of local technology, but international competition and sustainable development still need to be continuously explored.

Keywords: LLMs; Optimization Technology; Application Field; Privacy and Ethical Issues

1. Introduction

Since the Turing Test was proposed in the 1950s, language intelligence has remained a central research focus in artificial intelligence. Recent advances in deep learning and data-driven natural language processing have fundamentally reshaped this field.

After the evolution process from statistical model to neural model, the language modeling technology has developed into a key technical path to study the language understanding and generation mechanism. Large-scale pre-trained language model (PLMs) based on Transformer architecture shows significant NLP task generalization ability through massive corpus training (Shanahan, 2024). Empirical research shows that the model performance is significantly and positively correlated with the parameter size, and this phenomenon has triggered the theoretical discussion of the "parameter scale effect" in the academic circle (Zhao, 2023). As a milestone, large-scale language models not only achieve exponential performance improvements, but also show emergent capabilities such as context learning. Based on this, the term "Large Language Model"(LLMs) was formally proposed to define such PLMs with super-parameter scale (Wu et al., 2024). Its breakthrough in language modeling theory and NLP application practice opens up a new theoretical framework and technical path for the research of artificial intelligence (Naveed et al., 2023).

With the rapid iterative upgrade of generative pre-training models (such as GPT series), LLMs have triggered a global research and development boom and technological innovation (Wei et al., 2022). Using the GPT model developed by OpenAI, LLMs continuously refresh performance records in multimodal NLP benchmarks. Driven by the expansion of training data scale, driven by the innovation of algorithm architecture and the improvement of model complexity, such technologies have successfully penetrated into many vertical fields such as education informatization, intelligent government, fintech and biological medicine. This is supported by a bibliometric analysis of nearly 5,000 papers from 1996 to 2024, which reveals a shift in AI applications in e-commerce from purely technical algorithms to customer-centric applications such as personalized recommendations (Van et al., 2025). Similarly, in the field of e-commerce, empirical studies indicate that AI-based chatbots significantly enhance customer experience by improving responsiveness, ease of use, and personalization (Megdadi et al., 2025). Specifically in the public sector, systematic reviews indicate that LLMs significantly enhance government document management efficiency and classification accuracy when fine-tuned for domain-specific tasks, provided that challenges regarding legacy system integration and data governance are addressed (Yang et al., 2025). The current research progress in the field of artificial intelligence shows that the emergence ability of LLMs not only verifies the potential of AI technology, but also marks a revolutionary change in the paradigm of language intelligence research (Zhao et al., 2024).

Model training poses a significant technical barrier to the extreme demands of computing resources (such as the energy consumption cost of a GPU cluster). As LLMs expand, the computational resources required for their training increase exponentially, which not only leads to high research and development costs, but also brings serious environmental burden. The OpenAI The GPT-4 training process consumes about 25 million GPU hours and costs about $100 million. This energy-intensive model has also raised questions about the sustainability of AI technology (Kalyan, 2024).

In early 2025, the advent of DeepSeek has had a profound impact on the AI industry, especially in lowering the barriers to entry for the use of large models. Its success has prompted other AI companies to consider improving their models and could spark a new round of price competition (Sapkota et al., 2025). In code generation tasks, DeepSeek models perform well, even close to or beyond the top models such as GPT-4 Turbo. DeepSeek Technological innovation also challenges the traditional view that a lot of computing resources are needed to achieve top AI performance. DeepSeek Large models represent an important milestone in AI technology progress in China and around the world, demonstrating how to achieve high-performance AI solutions at a lower cost.

However, this field still faces multiple challenges: the black-box nature of LLMs leads to the lack

of interpretability of their working mechanism, which restricts the deepening of theoretical research (Yan et al., 2024). This "black box" feature makes it difficult to understand and verify the internal mechanism and logic of LLMs in the decision-making process, thus causing a crisis of confidence. For example, in high-risk areas such as health care and finance, users and regulators need to clearly understand the decision-making basis of the model, but the opacity of LLMs makes this goal difficult to achieve. Although the researchers have proposed a variety of methods to improve the interpretability of models, such as visualization of attention mechanism, LIME (locally interpretable model) and TransformerVis, these methods still have limitations to fully solve the black box problem. control of generated content has become a key problem restricting the implementation of technology (Huang et al., 2024). LLMs may generate bias, discrimination, or inaccurate information when generating text, which not only affects the user experience, but may also lead to social problems. LLMs may leak privacy or spread false information when dealing with sensitive topics (Xu et al., 2024). LLMs lack a sense of moral responsibility and in some cases may even make dangerous or improper judgments. Therefore, how to ensure that the output of LLMs meets the ethical standards through technical means has become an important direction of current research.

Despite advances in interpretability techniques, the black-box nature of LLMs remains unresolved, particularly in high-risk domains such as healthcare and finance, where ethical accountability and privacy protection are essential.

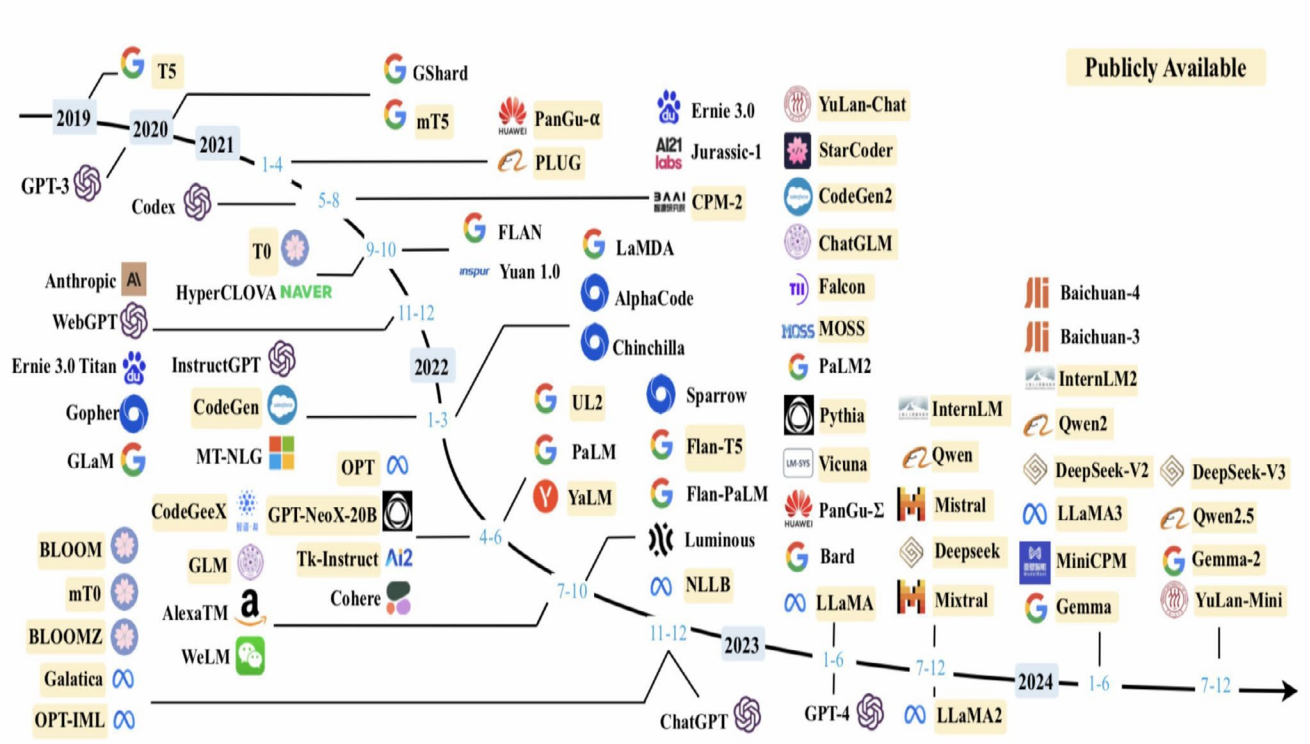

Fig. 1: The evolutionary path of LLMs in recent years

2. Literature Review

The earliest stage of large language model is actually natural language processing (NLP). The development process of natural language processing (NLP) can be divided into four main stages (Hadi et al., 2023). The technological breakthrough and paradigm shift in each stage significantly promote the exponential improvement of domain performance:

To improve structural balance, this section now focuses on representative milestones in LLM evolution rather than exhaustive historical narration, while analytical emphasis is shifted to optimization techniques and applications.

Early research in the rule-driven phase (1950s-1960s) focused on machine translation and basic

grammar parsing tasks, typical of the construction of artificial rule systems (such as dictionary mapping and syntactic template matching). Limited by the limitation of symbolism method, it is difficult to effectively deal with the complexity and diversity of natural language;

The jump in computing power in the statistical learning stage (1980s-1990s) and the accumulation of large-scale text corpus have led to the popularization of statistical methods. Among them, statistical machine translation (Statistical Machine Translation, SMT) based on bilingual parallel corpus realizes cross-language alignment through probabilistic modeling, and becomes the representative technology of this period. This stage lays the foundation of the data-driven paradigm, but it still faces the challenge of the complexity of feature engineering and the insufficient depth of semantic understanding;

Deep Learning Revolution (2010s-2020s) The innovation of deep neural network architecture (such as recurrent neural network, attention mechanism) and the proposal of Transformer model have completely reconstructed the NLP technology system. Through end-to-end distributed representation learning, the model has made breakthroughs in text understanding generation and cross-language transformation, and a number of benchmark performance exceeds traditional methods;

Multimodal fusion stage (2020s-) With the development of reinforcement learning framework and cross-modal alignment technology, the research focus is gradually extended to the collaborative modeling of multi-source heterogeneous data. The current work is dedicated to building a unified representational space integrating textual, visual, and auditory signals to achieve a semantic understanding closer to human cognition(Myers et al., 2024).

When discussing the development process of large language models (LLMs), the increase in the number of parameters is widely considered as one of the key factors to improve the model performance. The larger model capacity allows it to capture richer language information and contextual details, thus showing stronger capabilities in language understanding, generation, and transfer learning. As the size of the model grows, its ability to deal with complex language tasks and large-scale corpora has been significantly enhanced from early small models to today giant models with hundreds of billions of parameters. Take the OpenAI GPT series, for example:

As the first work of this series, GPT-1 aims to provide a solution for the single-sequence text generation task. Although it introduced a decoder structure based on the Transformer architecture at the time (12-layer one-way Transformer, Focusing only on the current and prior context) and innovative elements such as the pre-training-fine-tuning paradigm (Meyer et al., 2023).

But since the number of participants was only 117 million, Limiting their language understanding and generation capabilities, especially in the face of complex language tasks (such as long text logical reasoning) and large-scale corpora.

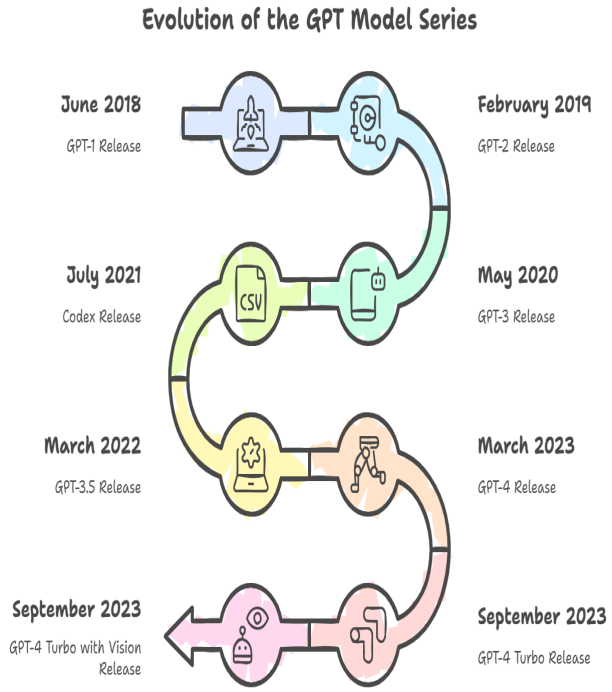

Fig. 2: ChatGPT development version and timeline

GPT-1 established the Transformer-based pre-training paradigm, enabling subsequent GPT-series models to achieve performance gains primarily through parameter scaling and data expansion rather than fundamental architectural changes.

The subsequent GPT-2 further improved the generalization and context processing capabilities of the model by significantly increasing the number of parameters to 1.5 billion and expanding the data set to 40GB. GPT-2 can not only generate more smooth and natural text, but also perform some simple reasoning tasks, but its potential abuse risks, such as generating fake news. Led the OpenAI to choose not to publicly release the full version. The multi-task learning strategy of GPT-2 demonstrates its powerful capabilities in natural language processing tasks.

In GPT-3, the parameter scale jumped to 175 billion, and the amount of training data reached hundreds of TB, which marked that LLMs have entered the era of super-large scale. GPT-3 not only demonstrates powerful language generation capabilities, but also is able to perform a variety of tasks, including translation, abstract, question and answer, without additional fine-tuning to a specific task (Liu et al., 2023). Due to the large parameter scale, the training costs and computational resources demand also surge, which makes the development and deployment of GPT-3 face a higher technical threshold. Google Is also actively exploring this area, with a series of models such as BERT, GlaM and LaMDA. The parameter size of LaMDA reached 137 billion, and the training sample size also reached 1.56TB. These advances not only reflect the technological progress, but also provide a valuable reference for subsequent studies.

Compared with its predecessor, GPT-4 represents a substantial leap in multimodal understanding and reasoning capabilities. Although OpenAI has not officially disclosed the exact parameter count, the model adopts a more advanced architecture trained on a massive scale of data. According to the GPT-4 Technical Report, it demonstrates human-level performance on various professional and academic benchmarks. For instance, GPT-4 scores in the top 10% of test-takers on a simulated Uniform Bar Exam and achieves state-of-the-art results on the MMLU (Massive Multitask Language Understanding)

benchmark, significantly surpassing previous models in complex reasoning and reliability tasks without relying on unverified parameter statistics.

With the success of ChatGPT and the global focus, the development of large models has entered a new stage. The number of large models in China shows an explosive growth trend, marking the rapid progress of domestic AI technology in this field. According to the 2025 China AI Model Market Scale and Product Analysis Report, by the end of 2024, the Cyberspace Administration of China has filed 302 generative AI services, of which 238 new ones were filed in 2024. Table 1 briefly presents several popular large models in China.

Table. 1: Introduction to the popular big models in China

Model Name	Company	Model Overview
Wenxin Model	Baidu	A comprehensive model integrating NLP and cross-modal generation; widely deployed in search engines, content creation, and enterprise cloud solutions.
Tongyi Model	Alibaba	Contains Tongyi Qianwen (text generation), Tongyi Wanshang (image generation), supporting enterprise intelligent upgrades.
Dòubāo Model	ByteDance	Lightweight model focusing on efficient inference and low-cost deployment
Hunyuan Model	Tencent	Mixture of Experts (MoE) model, excelling in Chinese evaluations with significantly improved inference efficiency .
Pangu Model	Huawei	Industry-level model series covering mining, meteorology, finance, and other verticals, empowering digital transformation of industries.
DeepSeek Model	DeepSeek	Open-source model that has shown outstanding performance in code generation in recent updates.

Rather than providing product-level descriptions, this subsection synthesizes common architectural trends, deployment strategies, and efficiency-oriented optimization paths among representative Chinese LLMs.

For example, the Wenxin big model: developed by Baidu, Covering capabilities such as natural language processing, cross-modal generation, Widely used in search, content creation and other scenarios; Tongyi Big model: a multimodal large model launched by Alibaba, Including questions (text generation) and universal (image generation), Support enterprise intelligent upgrading; Spark big model: The cognitive intelligence big model developed by iFlytek,Focus on education, health care and other fields, Multi-lingual interaction ability; Big model: a lightweight model introduced by ByteDance, Focus on efficient reasoning and low-cost deployment, Serving in TikTok and other ecological applications; Mixed yuan big model: Mixed expert model (MoE) developed by Tencent (Liang et al., 2024), Excellent performance in the Chinese assessment, The reasoning efficiency is significantly improved compared with the previous generation; Pangu grand model: an industry-level model series launched by Huawei, Covering vertical fields such as mining, meteorology and finance, Digital transformation of the enabling industry; Wisdom spectrum Qingyan: a general dialogue model developed by Wisdom spectrum Huazhang, Based on the large-scale and high-quality data training, Support for complex logical reasoning; Vidu video large model: jointly developed by Shengshu Technology and Tsinghua University, China's first long-time video generation model for benchmarking Sora, Realize high dynamic content creation; DeepSeek Big model: an open source model launched by Deep Search (DeepSeek), The recent upgrade has performed well in code generation and mathematical

reasoning tasks; ChatGLM Series: A bilingual model jointly developed by Wisdom Music Chapter and Tsinghua University, The open-source versions are widely used in both academia and industry. Since 2024, these large models have not only increased in number, but also made great progress in technological innovation and application scenario expansion.

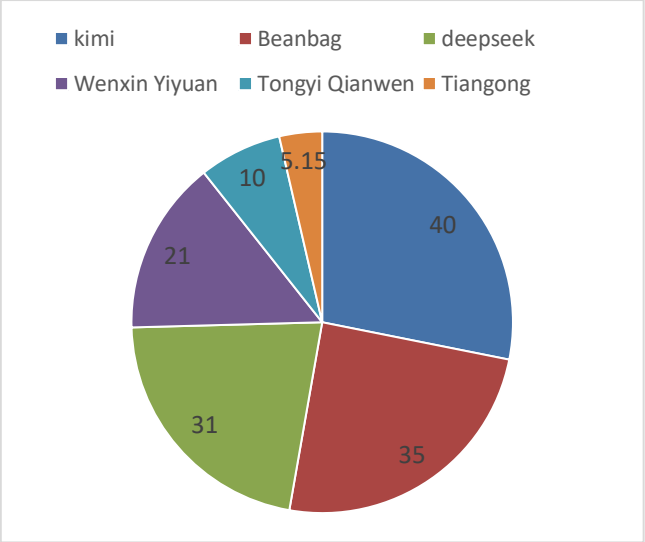

Fig. 3: Statistics of the number of popular big models used in China

Statistically, the website traffic, plus the total downloads of the Android mobile app.similarweb, Query the website traffic; appbrain, query the total download of Android app; Tencent App Bao: query the download of Chinese app app. App downloads of Apple users were not counted because they could not get data.

Kimi big model is an AI tool launched by Beijing Yuezhi Dark Face Technology Co., LTD. Its core advantage is its excellent long text processing ability. Kimi supports lossless context processing of up to 2 million words, which is at the leading level in the world. It can effectively solve the problem of "fragments" of traditional large models when dealing with long texts, enabling it to perform well in complex scenarios such as legal document analysis, academic paper summary, and market analysis. Kimi also has a powerful memory function, which can maintain the integrity and continuity of the dialogue, while supporting multiple files and the resolution of multiple formats, such as TXT, PDF, Word, etc., which greatly facilitates users' data sorting and information query. Kimi also has some shortcomings. Despite its strong long text processing capabilities, some users report that it has a high failure rate in translating specific content, and have limited ability to identify structured information and sloppy text. Kimi has not yet achieved perfection in generating continuous conversations and generative capabilities, especially poorly in prosody and numerical accuracy (Raiaan et al., 2024). Kimi's training data is limited, leading to its generalisability and multilingual support, especially in non-Chinese environments with less performance than international large models. And Kimi does not support private deployments, which may limit the needs of some enterprise users.

Through the deep upgrade of self-developed machine learning framework Angel, Tencent mixed yuan large model has achieved multi-dimensional breakthroughs in the aspects of training efficiency, storage capacity and computing power cost optimization. AngelPTM Training framework breakthrough, training efficiency reaches 2.6 times of mainstream open source frameworks (such as PyTorch, DeepSpeed), 100 billion level model training can save 50% of computing power cost. Through the storage optimization of 90% video memory capacity expansion, combined with multi-dimensional parallel technology (data parallel, model parallel, flow parallel, sequence parallel), the bottleneck of video memory is significantly alleviated. Communication optimization using hardware and software collaboration, including bandwidth broadening, GPU topology sensing and load balancing technology, is adopted to improve the efficiency of distributed training. Develop heterogeneous operator

compilation layer, compatible with domestic chip architectures (such as Shengteng and Haiguang), support 10,000 card-level super-scale training, and enable Tencent Cloud HCC large model exclusive computing power cluster. The world is the first in adopting the MoE architecture. Through Expert computing and communication overlap optimization, operator fusion and other technologies, the model effect is improved by 50%, and the training performance exceeds the DeepSpeed framework by 2.6 times. Supporting open source AngelHCF-vLLM reasoning framework, low precision training optimization and domestic hardware adaptation. Using a 3-layer fully self-developed network architecture (switch, optical module, network card), the single cluster supports a scale of 128,000 cards and a communication bandwidth of 3.2T. Improve the GPU utilization rate by 40%, reduce the training cost by 30% -60%, and improve the communication performance by 10 times. Inference optimization: By extending parallel capabilities, operator optimization and batch processing technology, the reasoning speed is 1.3 times higher than the mainstream framework in the industry, and the cost is reduced by 70%. For example, the time of the graphic scene is reduced from 10 seconds to 3-4 seconds. One-stop platform support: covering the whole process of model pre-training, fine tuning, reinforcement learning, support vector database, automatic fine-tuning and other tools, and has served more than 300 internal business scenarios such as Tencent conference and news. Large model technology system ranked the first echelon in China in the 2024 SuperCLUE evaluation, and its relevant achievements won the first prize of Science and Technology of China Institute of Electronics in 2023, highlighting the leadership of full-link self-research technology.

The V3.0 version of the Spark cognitive model is released and claims that the medical field capability exceeds GPT-4, reflecting the depth optimization and development of the application of domestic large model in specific fields. V3.0 improves the seven core abilities (text generation, language understanding, knowledge quiz, logical reasoning, mathematical ability, code ability, multi-modal ability), and has comprehensively surpassed ChatGPT (based on GPT-3.5) in the Chinese understanding evaluation, and its English ability is comparable to ChatGPT. Especially in the field of medical treatment, model for massive knowledge questions, complex language understanding, professional text generation and diagnosis and treatment recommended scenarios for special optimization, national comprehensive utilization of science and technology information resources and public service center (STI) of the third party test shows that the actual use data in 120000 sampling questions beyond the GPT-4.

As a new generation of large language model launched by Ali Cloud, Tongyi Qianwen version 2.5 has achieved an all-round breakthrough in the parameter scale, technical capability and application landing level. According to the OpenCompass benchmark test, Tongyi QianQ 2.5 scores GPT-4 Turbo, which is the first time that the domestic large model has achieved the first place tied with the top closed source model in this authoritative evaluation. Specifically, its understanding ability has improved by 9%, logical reasoning by 16%, instruction compliance by 19%, and code ability by 10%. In the Chinese context, knowledge answering,

text generation and security risk control have surpassed GPT-4. Long document processing: support a single input of 10 million words, 100 documents, Covering various formats such as PDF, Word, and Excel, Accurate extraction of titles, paragraphs, charts and other structured information; Multimodal upgrade: significantly enhanced audio and video understanding ability, Tongyi Lingcode Enterprise Edition is based on CodeQwen1.5 technology, Plugoutins were downloaded more than 3.5 million times; Open source ecological construction: release the Qwen1.5-110B open source model with 110 billion parameters, Llama-3-70B beyond Meta in benchmarks like MMLU, TheoremQA, On the top of the Hugging Face open source big model list.

In July 2023, The team established Hangzhou Deep Search (DeepSeek) Artificial Intelligence Basic Technology Research Co., LTD., focusing on the research and development of general artificial intelligence (AGI) and large models.The initial fund is supported by The quantitative 5. On November 2 of the same year, DeepSeek released DeepSeek Coder, which was its first open source code large

model released in the AI field, marking the company's breakthrough in the field of code intelligence (Guo et al., 2024). On January 5,2024, DeepSeek launched its first large-scale language model, DeepSeek LLM, which has reached 67 billion parameters, and uses optimized learning rate schedulers and innovative alignment techniques to perform well in language understanding and generation tasks. In May 2024, DeepSeek-V2 was released as a second-generation hybrid expert (MoE) large model, with 236 billion parameters, introducing the long-head potential attention (MLA) mechanism, which significantly improved the reasoning efficiency and the ability to deal with complextasks. On December 26,2024, DeepSeek-V3 was released. This third-generation general large model further improves performance, using cutting-edge technologies such as non-auxiliary loss load balancing strategy and multi-word element prediction (MTP), to greatly shorten the time of content generation.

On January 20,2025, DeepSeek released the open-source inference model DeepSeek-R1, which achieved comparable performance to OpenAI o1 in inference tasks and attracted wide attention with extremely low training cost. In February 2025, DeepSeek released five core tools, including FlashMLA, DeepEP, DeepGEMM, DualPipe and Optimized Parallelism Strategies, which brought a qualitative leap in improving the efficiency of model training and reasoning. The DeepSeek-R1, an important version of the DeepSeek, has attracted the attention of the global tech community when it was released. With its efficient performance and relatively low cost, the model has near or outperformed competitors in multiple benchmarks. The DeepSeek-R1 performed well in the mathematical reasoning task. In testing on the MATH dataset, DeepSeek-R1 solved many complex problems that were previously inaccessible by the model with a multi-step inference approach, with significantly better accuracy than other token-based models. The DeepSeek-R1 also demonstrates its powerful ability in handling complex tasks, especially in applications of reinforcement learning. By optimizing the reward function and improving the data processing strategy, the model can reason more efficiently with small amounts of annotated data (Choi & Chang, 2025). The efficient performance of DeepSeek-R1 also benefits from its innovative approach during training. Group Relative Strategy optimization (GRPO) is introduced to reduce the training cost of reinforcement learning, and human preference alignment techniques are combined to further improve the accuracy and generalization ability of the model.

This training strategy not only improves the efficiency of the model, but also makes it excellent in multi-task learning, such as continuous improvement in math and encoding tasks. Although DeepSeek-R1 shows excellent performance in several fields, it still has limitations in certain application scenarios. In tasks that require rapid response, efficiency may be affected due to its reliance on generating large numbers of tokens for inference. Therefore, future studies can further optimize the balance between their inference speed and accuracy to meet more practical application needs.

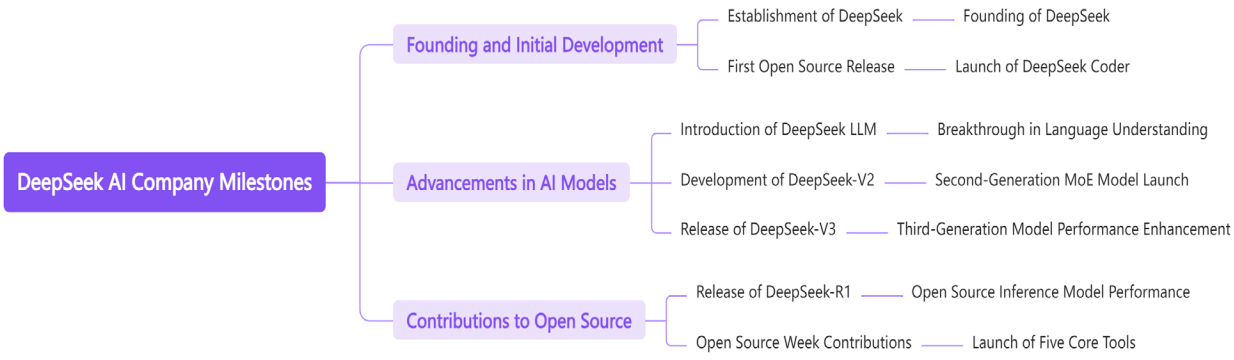

Fig. 4: DeepSeek development version

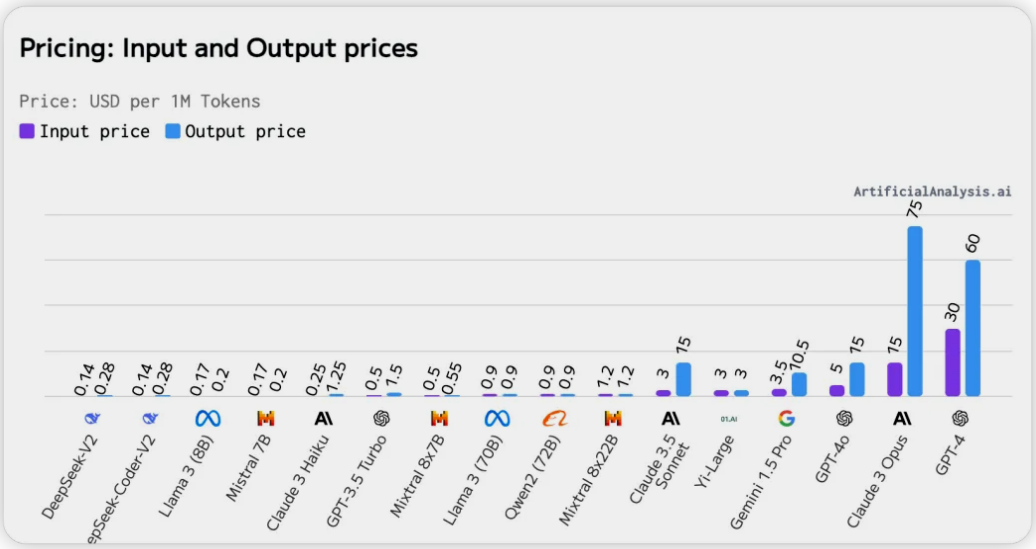

Fig. 5: Input and Output prices

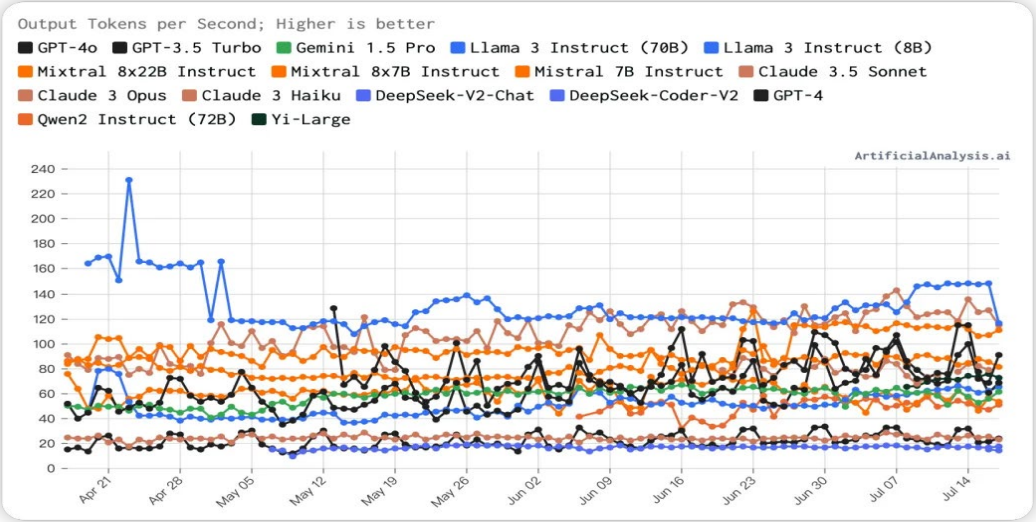

Fig. 6: Output tokens per second

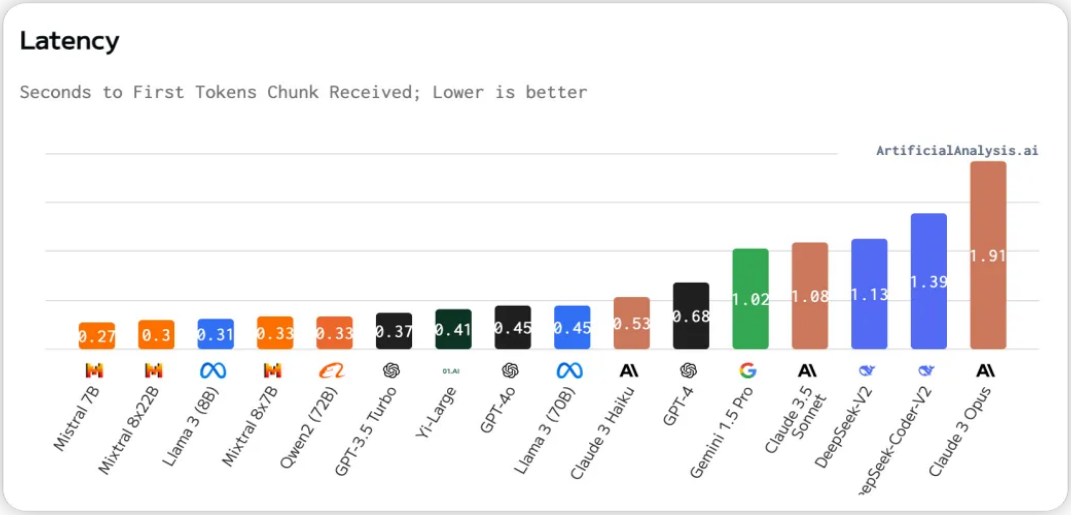

Fig. 7: Delays for different models

3. Methodology

3.1.Research Design

This study adopts a hybrid research framework combining a Systematic Literature Review (SLR) with a quantitative comparative analysis to comprehensively evaluate the evolutionary trajectory and technological landscape of Large Language Models (LLMs). By integrating historical retrospection with empirical performance data, the research moves beyond a mere descriptive summary to structurally analyze the "parameter scale effect" and the paradigm shift from rule-driven NLP to Transformer-based architectures. The design specifically focuses on bridging the gap between theoretical optimization techniques—such as pre-training and prompt learning —and their practical deployment in vertical domains, while simultaneously conducting a case study on the rise of indigenous Chinese models like DeepSeek to assess current competitive dynamics in the global AI market.

3.2.Data Collection and Sources

To ensure a robust evidence base, data collection was bifurcated into academic literature and industrial performance metrics, covering a temporal range from early NLP developments to the contemporary era. For the bibliometric analysis, we accessed core databases including Web of Science and Google Scholar to retrieve approximately 5,000 relevant papers spanning from 1996 to 2024, utilizing keywords such as "Large Language Models," "Transformer Architecture," and "Generative AI Applications." This academic data is supplemented by authoritative industry reports, such as the 2025 China AI Model Market Scale and Product Analysis Report, and technical benchmarks from OpenCompass and SuperCLUE , which provide real-time data on model parameters, training costs, and market filings of generative AI services.

3.3.Inclusion and Exclusion Criteria

The selection process prioritized peer-reviewed articles and authoritative technical reports that offer substantive insights into model architecture, optimization algorithms, and performance evaluations, while filtering out purely news-based or redundant content. Inclusion criteria were strictly defined to encompass studies detailing the technical evolution from statistical models to deep neural networks, research verifying emergent abilities and generalization in super-parameter scale models, and specific technical documentation regarding the optimization and cost-efficiency of recent Chinese models like DeepSeek-V3 and DeepSeek-R1. Conversely, literature lacking empirical validation or focused solely on outdated rule-based systems without reference to modern deep learning integration was excluded to maintain the review's relevance to the current state-of-the-art.

3.4.Comparative Analysis Framework

The analytical framework employs a multi-dimensional metric system to evaluate model efficacy, specifically contrasting international baselines (e.g., GPT-4) with emerging Chinese iterations (e.g., DeepSeek, Tongyi). This comparison is conducted through three primary lenses: technical specifications, including parameter scale and architectural innovations like Mixture-of-Experts (MoE); economic efficiency, analyzing quantitative indicators such as training GPU hours, input/output pricing per million tokens, and inference latency; and domain adaptability, assessing performance in specific tasks like mathematical reasoning, coding, and long-text processing. This structured comparison aims to validate the hypothesis that optimization techniques can achieve high-performance AI solutions at significantly lower computational costs.

4. Results and Discussion

4.1.Optimization technique

There are many optimization techniques for training large language models, which can be roughly divided into three categories: pre-training, prompt learning, and model fine-tuning(Sahoo et al., 2024).

This section integrates optimization techniques with application requirements by mapping pre-training, prompt learning, and fine-tuning strategies to different service contexts, highlighting trade-offs among performance, computational cost, and governance.

4.1.1. Pre-training technology

The pre-training model uses self-supervised learning to extract the common language features from massive unannotated data, a technical paradigm that started with the Word2Vec model in 2013. Word2Vec Generate static word vectors by jump-gram and CBOW architecture. However, its limitation is the inability to deal with the problem of polysemy, such as the semantic difference of "apple" in the context of fruit and technology companies is completely ignored. In order to break through this bottleneck, ELMo in 2018 adopted a two-way LSTM architecture to learn context-related word vectors through hierarchical superimposed language models, which significantly improved the ability to capture semantic and syntactic features.

The real technological revolution stems from breakthroughs in the Transformer architecture. The BERT model proposed in 2018 completely abandons the recurrent neural network, uses multi-layer Transformer encoder stacking, and realizes deep two-way representation learning through mask language modeling (MM) and the next sentence prediction (NSP) task. The MLM strategy randomly covers 15% of the input word elements, forcing the model to reconstruct the masked content from a two-way context. This self-supervision mechanism enables BERT to capture the different semantics of "cell" in biological and communication scenarios. With a 768-dimensional hidden layer and 12 attention heads configuration, BERT-base scored a breakthrough score of 80.5 on the GLUE benchmark. Subsequent optimization focused on training strategy innovation: RoBERTa Empirical studies with dynamic masks, larger batches (8k tokens), and longer training periods (500k steps) showed that the original BERT has only completed about 55% of the theoretical training volume.

It is worth noting that the joint pre-training of multi-language BERT (mBERT) in 104 languages proves that cross-language transfer learning has a significant effect on low-resource languages, and only 1% of the target language annotation data can achieve the full data performance of traditional methods. Tip learning (Prompt Learning) is a kind of rapid development in the field of natural language processing (NLP) in recent years, its core idea is by adding "prompt" information in the input data, adjust the form of the input data, make it closer to the pre-training stage of data form, to reduce the difference between pre-training and fine-tuning stage data form, make the model can be efficiently used for downstream tasks. The goal of cue learning is to improve the performance of the model in downstream tasks by optimizing the design of cue information, while reducing the need for large amounts of annotated data and complex fine-tuning.

4.1.2. Tips for learning technology

The cue learning techniques can be divided into discrete prompts and continuous prompts according to their form.

Discrete prompt often consists of natural language texts, such as artificially designed templates or phrases found automatically by search algorithms, characterized by each word being independent and not constrained by the word vector. The research of discrete prompts is mainly divided into two routes: one is to use automatic search and build optimal discrete prompts, such as gradient search, bundle search, etc.; the other is to manually design the form of discrete prompts to stimulate the ability of large language models, including context learning and thought chain. Context learning allows the model to be solved directly in a few-sample natural language processing task without updating any model parameters; the thought chain refers to the ability of the model to generate inference process and generate answers in the reasoning task, which can be achieved through few-sample or zero-sample setting.

Continuous prompt refers to the prompt information is a continuous vector not constrained by the word vector, and usually contains some trainable parameters, which can be updated in the training of

downstream tasks(Thistleton & Rand, 2024). Compared to discrete prompts, continuous prompts have higher flexibility and adaptability that can better capture the requirements of complex tasks. Continuous prompts can be optimized by prefix tuning (Prefix Tuning), P-tuning and other methods, which improve the performance of the model on downstream tasks by adding a specific task-related vector sequence before the input sequence prompt learning technology column, while keeping the pre-trained model parameters unchanged. Moreover, continuous prompts can also embed discrete prompts into continuous space through soft linear combination, further enhancing the expression ability of the model.

Fig. 8: Optimization technique route

4.1.3. Model fine-tuning technology

The core idea of model fine-tuning techniques is to use a language model pretrained on large-scale text data and further trained on data for specific tasks or fields to accommodate specific tasks. The procedure involves initializing the model using the parameters of the pre-trained model, followed by training on the task-specific dataset. This dataset typically includes user interactions with projects, project descriptions, user information, and other relevant contextual data. In sentiment analysis, named entity recognition, quiz system and other tasks, fine-tuning can significantly improve the performance of the model. The implementation of fine-tuning technology usually includes the following steps: first, select the appropriate pre-trained model as the basis, such as BERT, GPT, etc; prepare the annotated task data set; then adjust the model parameters to better adapt to the target task; and finally evaluate and optimize the model. It is worth noting that the hyperparameters, such as learning rate, regularization strategy, need to be reasonably set up in the fine-tuning process to avoid overfitting.

In recent years, fine-tuning technology has made breakthroughs in several fields. In social media analysis, BERT model can effectively capture user emotional and behavior patterns through fine tuning; in the medical field, BioBERT achieves efficient processing of biomedical data through fine tuning; and in education, BERT-based model improves the performance of text classification and question answering system. The fine-tuning technique not only reduces the cost of model training, but also improves the generalization ability of the model to better adapt to the needs of specific tasks. In the future, with the continuous development of pre-training models and fine-tuning technologies, its application prospects in natural language processing and other fields will be even broader.

4.2.Application area

As the ability of the big language model becomes more and more powerful, its role in various fields is also increasing.Across education, finance, healthcare, and manufacturing, LLM applications share common patterns such as decision support and service automation, while facing shared challenges in reliability, domain adaptation, and data governance.

4.2.1. Education

The research status and application prospect of large language models (LLMs)in the field of education show great potential and challenges. A systematic review covering literature from 2017 to 2025 further confirms this trend, highlighting the role of conversational AI in providing scalable and personalized

learning solutions while noting persistent challenges such as lack of empathy and technical limitations (Alkishri et al., 2025). From the perspective of teachers, LLMs can significantly improve the teaching efficiency, and provide teachers with convenient teaching AIDS and resources. For example, LLMs can help teachers in curriculum planning, generate syllabi and curriculum plans through intelligent dialogue systems, and quickly identify key points and difficult points in teaching, so as to optimize teaching design (Wang et al., 2024). LLMs can also support differentiated and personalized teaching, providing customized learning materials and feedback according to students' learning progress and needs, to help teachers better meet the needs of different students. In assessment, LLMs can automatically correct answers to open questions, saving teachers time and enabling them to focus on more creative teaching methods (Yan et al., 2024). At the same time, LLMs can also support teacher professional development and promote collaboration and growth among teachers by providing educational resources and sharing teaching methods.

For students, LLMs are of great value in improving their learning experience and abilities. LLMs are able to generate personalized learning content, including reading materials, exercises and explanations, to help students consolidate knowledge and improve learning efficiency. In math learning, LLMs can help students overcome difficult problems through conversational interactions and design adaptation tests to accurately assess students' knowledge level (Bewersdorff et al., 2025). LLMs are also able to provide real-time feedback and guidance, enhancing students' sense of participation and motivation to learn. Relevant research further points out that attitude plays a key mediating role in shaping students' behavioral intention to adopt and utilize generative AI tools like ChatGPT(Paudel & Acharya, 2024).Furthermore, comparative experiments in university student support systems demonstrate that generative models like GPT-4 significantly outperform traditional methods in accuracy and empathy dimensions, validating their potential for enhancing service quality, though they require human oversight to mitigate hallucination (Eirena & Shah, 2025a)s. In language learning, LLMs can simulate natural dialogue and help students improve their language ability. LLMs also face several challenges, such as data privacy and security issues. Since LLMs need to deal with a large amount of user data, how to protect students' privacy and data security has become an urgent problem in the education field. To address this, studies suggest that a hybrid Retrieval-Augmented Generation (RAG) framework combining locally deployed LLMs with semantic vector search can effectively ensure data sovereignty and privacy security while maintaining high real-time response performance (Eirena & Shah, 2025b).

4.2.2. Finance

The application of LLMs in the financial field is not limited to prediction and analysis, but also includes risk management, customer service automation, fraud detection and compliance review. LLMs can identify potential compliance risks and provide actionable advice by parsing complex regulatory documents. In terms of customer service, LLMs can provide real-time and personalized service to customers through natural language processing technology, thus improving customer satisfaction and loyalty (Lee et al., 2025). Regarding specific applications in banking, studies using the Extended Technology Acceptance Model (TAM) reveal the critical factors influencing Generation Z's adoption of such AI services, highlighting the importance of technical adaptability (Lim et al., 2025).

Although LLMs show great potential in the financial field, their application still faces many challenges. The high specialization and dynamics of financial data requires the high adaptability of the model. LLMs need to ensure the accuracy and reliability of the output when dealing with high-risk decisions, which requires the combination of expert system and manual audit mechanism to further enhance the credibility of the model. To overcome these challenges, the researchers propose multiple solutions. With instruction tuning (instruction fine-tuning) and alignment tuning (alignment fine-tuning), LLMs can better understand professional issues in the financial field and generate outputs that align with human values or preferences. Multimodal models combined with traditional machine learning methods have also been shown to improve the accuracy and efficiency of financial prediction (Xie et

al., 2024). Due to the high complexity and accuracy requirements of this field, scholars are working on the development of financial-specific LLMs with high accuracy and high efficiency. Future research directions include further optimizing the model's adaptability, improving its reliability in high-risk decisions, and exploring more efficient fine-tuning methods.

4.2.3. medical treatment

The application of large language model (LLM) in the medical field has made remarkable progress in recent years, and its potential and practical effects have gradually been widely recognized. Through deep learning and large-scale data training, these models can perform a variety of complex tasks, thus bringing innovative changes to the healthcare industry (He et al., 2025).

In terms of clinical diagnosis, large language models are able to assist doctors in disease prediction, diagnosis and treatment plan formulation. They can be made by analyzing the patient's symptoms, history and medical literature, generate personalized diagnostic advice and treatment plan, so as to improve the accuracy and efficiency of diagnosis based on multimodal technology model can combine a variety of data sources such as image, text, further enhance the accuracy of disease detection, such as radiology imaging analysis and pathological report generation (Alber et al., 2025).

In the field of drug development, large language models demonstrate powerful text processing capabilities. They can quickly analyze massive amounts of chemical and biological data, predict the binding strength of compounds and targets, and screen potential new drug candidates, thus accelerating the new drug development process and reducing research and development costs. These models are also able to generate intelligent writing content, translating complex medical information into easy-to-understand language and helping non-professionals acquire expertise (Labrak et al., 2024).

In patient care and education, large language models provide patient services such as health counseling, postoperative follow-up and psychological support through natural language generation techniques. For example, they can simulate the expression of doctors, providing personalized health education and psychological counseling for patients, thus improving the patient's medical experience. These models also help healthcare workers complete paperwork efficiently, such as medical record writing and surgical reports.

Large-scale language models also play an important role in medical management and operations. They can automate cumbersome tasks such as medical billing and medical record coding, reduce human error and improve work efficiency. In addition, by building knowledge maps and analyzing medical data, these models can also improve the optimal allocation of medical resources and improve the operational efficiency of medical institutions. Although large-scale language models show broad application prospects in the medical field, their development still faces many challenges. The quality and diversity of data required for model training directly affect its performance; ethical, privacy, and legal issues also need to be appreciated. With the continuous progress of technology and the improvement of relevant regulations, large-scale language models are expected to play a greater role in the medical field and promote the intelligent upgrading of the medical industry.

4.2.4. Manufacturing industry

The application of large language model (LLM) in the manufacturing field is becoming an important force in promoting intelligent manufacturing and industrial upgrading. In recent years, with the rapid development of artificial intelligence technology, LLM growing role in manufacturing, its core advantage lies in the strong natural language processing ability, deep learning ability and multimodal interaction ability, make it can for production process optimization, equipment maintenance, quality management, supply chain management, and other links to provide intelligent support (Baptista et al., 2025).

In terms of production process optimization, LLM is able to predict production bottlenecks and optimize production planning and scheduling by analyzing historical and real-time data, thus significantly improving production efficiency (Fan et al., 2024). By combining machine learning

algorithms, LLM can analyze equipment usage data, predict equipment maintenance requirements, reduce downtime while optimizing the operational efficiency of production lines. LLM can also combine process literature and technical standards to improve the process, such as adopting new technologies or changing the use of materials in key processes, to further improve production efficiency and product quality. In terms of equipment maintenance and fault prediction, LLM can realize fault prediction and optimize the maintenance plan through the analysis of operational data and historical maintenance records. In semiconductor manufacturing, LLM combined with CIM 2.0 system can break data islands, optimize the use of tools, and improve equipment operation stability and capacity. This intelligent maintenance method not only reduces the maintenance cost, but also improves the service life and reliability of the equipment. In the field of quality management and detection, LLM combined with image recognition technology can greatly improve the speed and accuracy of automatic defect detection. For example, on the production line, LLM can analyze image data in real time, quickly identify product defects, and generate feedback reports, thus reducing manual intervention and improving detection efficiency. In terms of supply chain management, LLM optimizes inventory management and logistics processes by integrating and analyzing large amounts of data, and improves the elasticity and response speed of the supply chain (Liu et al., 2024). For example, in high-end manufacturing, LLM can analyze global supply chain data to optimize material use, production costs, and delivery times in real time. This capability enables companies to respond more flexibly to market changes and improve their competitiveness.

LLM also plays an important role in product design and optimization. By analyzing the insights and practical needs of frontline operators, LLM is able to generate designs that meet realistic conditions and make recommendations for sustainable materials and manufacturing processes. This not only improves design efficiency, but also drives innovation.

5. Conclusion

Large language models (LLMs)have achieved a revolutionary breakthrough in language intelligence through deep learning and massive data training. At the technical level, the evolution from statistical model to Transformer architecture, as well as optimization methods such as pre-training, prompt learning and fine-tuning, significantly improve the generalization ability and task adaptability of the model. Very large scale parameter models (such as GPT-4, DeepSeek) show multimodal understanding and emergence capabilities, but also expose problems such as high energy consumption, black box characteristics and ethical risks. In terms of applications, LLMs have been effective in education, finance, medical care and manufacturing, such as personalized teaching, financial prediction, clinical assisted diagnosis and intelligent manufacturing optimization. Chinas big model (such as Tongyi Qianwen and Flytek Spark) performs well in specific tasks, but it needs to further break through technical barriers and ecological construction.

Beyond summarizing existing surveys, this review contributes a structured linkage between optimization techniques and service-oriented application domains, and identifies research opportunities in logistics, informatics, and service science.

As the scale of large language models (LLMs)expands and the task complexity increases, it becomes particularly urgent to improve the interpretability, security, and ethical compatibility of the models. While combining LLMs with multimodal learning shows great potential, this process is accompanied by several challenges:

1. Computational resources and data problems: With the increase of model parameters and the diversification of training data, LLMs face challenges such as increased demand for computing resources, data bias and fairness, model interpretability and multi-modal fusion.

Future work needs to optimize the model framework and training methods, such as incremental learning algorithms, domain adaptation, multimodal representation learning, cross-modal alignment and fusion, and multimodal generation technologies.

2. Personalization and privacy protection: Personalized and customized LLMs technology is an important direction of future development, which will help to solve the problem of insufficient training data of large models.

How to protect user privacy and prevent the abuse of user data while providing personalized services will be a key problem. Further research on the application of key technologies such as differential privacy, federated learning, and personalized model fusion is needed to reduce the risk of data leakage.

3. Social ethics and governance: Accountability mechanisms such as auditing, impact assessment and certification help ensure that AI systems comply with ethical, legal and technical requirements. Existing audit procedures fail to fully address the governance challenges posed by LLMs,and challenges and complexity in implementing effective accountability remain. In the future, further research on key technologies such as ethical review, moral framework and legal compliance should be carried out to ensure the moral use and social responsibility of the large language model.

References

Alber, D. A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A. A., ... & Oermann, E. K. (2025). Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 1–9.

Alkishri, W., Yousif, J. H., Al Husaini, Y. N., & Al-Bahri, M. (2025). Conversational AI in education: A general review of chatbot technologies and challenges. *Journal of Logistics, Informatics and Service Science*, *12*(3), 264–282.

Baptista, M. L., Yue, N., Manjurul Islam, M. M., & Prendinger, H. (2025). Large language models (LLMs) for smart manufacturing and Industry X.0. In *Artificial Intelligence for Smart Manufacturing and Industry X.0* (pp. 97–119). Springer Nature Switzerland.

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., ... & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences*, *118*, 102601.

Choi, W. C., & Chang, C. I. (2025). Advantages and limitations of open-source versus commercial large language models (LLMs): A comparative study of DeepSeek and OpenAI's ChatGPT.

Eirena, A., & Shah, N. (2025a). Generative AI in university customer service: A comprehensive framework for enhancing efficiency and experience. *Journal of Logistics, Informatics and Service Science*, *12*(7), 56–67.

Eirena, A., & Shah, N. (2025b). A hybrid RAG-based chatbot for university customer service: Combining local LLM with semantic retrieval for privacy and real-time performance. *Journal of Logistics, Informatics and Service Science*, *12*(8), 58–72.

Fan, H., Liu, X., Fuh, J. Y. H., Lu, W. F., & Li, B. (2024). Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. *Journal of Intelligent Manufacturing*, 1–17.

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., ... & Liang, W. (2024). *DeepSeek-Coder: When the large language model meets programming—The rise of code intelligence*. arXiv.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). *A survey on large language models: Applications, challenges, limitations, and practical usage*. Authorea Preprints.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2025). A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *Information Fusion*, 102963.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., ... & Zhao, Y. (2024, July). Position: TrustLLM: Trustworthiness in large language models. In *International Conference on Machine Learning* (pp. 20166–20270). PMLR.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and applications of large language models*. arXiv.

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048.

Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10), 260.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024). *Biomistral: A collection of open-source pretrained large language models for medical domains*. arXiv.

Lee, J., Stevens, N., & Han, S. C. (2025). Large language models in finance (FinLLMs). *Neural Computing and Applications*, 1–15.

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., & Liu, K. (2024, January). A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering* (pp. 405–409).

Lim, K. B., Low, K. H., Yeo, S. F., & Tan, C. L. (2025). Factors influencing Generation Z's adoption of AI in banking: An extended technology acceptance model approach. *Journal of Logistics, Informatics and Service Science*, *12*(4), 178–192.

Liu, X., Yang, S., Dong, X., Rong, H., & Fu, B. (2024, September). Manu-Eval: A Chinese language understanding benchmark for manufacturing industry. In *China Conference on Knowledge Graph and Semantic Computing* (pp. 309–317). Springer Nature Singapore.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017.

Megdadi, O., Al-Ahmed, H., Ashour, M. L., Shriedeh, F. B., & Alshaketheep, K. (2025). The impact of chatbots on customer experience in e-commerce: Examining responsiveness, ease of use, and personalization. *Journal of Logistics, Informatics and Service Science*, *12*(7), 147–163.

Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P. C., ... & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, *16*(1), 20.

Myers, D., Mohawesh, R., Chellaboina, V. I., Sathvik, A. L., Venkatesh, P., Ho, Y. H., ... & Jararweh, Y. (2024). Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts. *Cluster Computing*, *27*(1), 1–26.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). *A comprehensive overview of large language models*. arXiv.

Paudel, S. R., & Acharya, N. (2024). Factors affecting behavioral intention to use ChatGPT: Mediating role of attitude. *Journal of Service, Innovation and Sustainable Development*, *5*(2), 143–162.

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839–26874.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A systematic survey of prompt engineering in large language models: Techniques and applications*. arXiv.

Sapkota, R., Raza, S., & Karkee, M. (2025). *Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SOTA large language models*. arXiv.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79.

Thistleton, E., & Rand, J. (2024). Investigating deceptive fairness attacks on large language models via prompt engineering.

Van, N. D., Hoang, C. C., & Khoa, B. T. (2025). Bibliometric examination of artificial intelligence within the framework of e-commerce technology from 1996 to 2024. *Journal of Logistics, Informatics and Service Science*, *12*(2), 138–150.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., ... & Wen, Q. (2024). *Large language models for education: A survey and outlook*. arXiv.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). *Emergent abilities of large language models*. arXiv.

Wu, T., Luo, L., Li, Y. F., Pan, S., Vu, T. T., & Haffari, G. (2024). *Continual learning for large language models: A survey*. arXiv.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., ... & Huang, J. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, *37*, 95716–95743.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). *Hallucination is inevitable: An innate limitation of large language models*. arXiv.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). *On protecting the data privacy of large language models (LLMs): A survey*. arXiv.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., ... & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112.

Yang, H., Nawi, H. S. A., & Zhang, Y. (2025). Artificial intelligence and large language models in government document management: A systematic review of applications, challenges, and implementation strategies. *Journal of Logistics, Informatics and Service Science*, *12*(4), 129–145.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). *A survey of large language models*. arXiv.