ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202310.00067

Technical Characteristics of Large Language Models and Recent Market Developments: A Postprint

Authors:

Date: 2023-10-08T00:00:00+00:00

Abstract

In recent years, large language model technology has advanced by leaps and bounds. The rapid technological development and expansion of internet enterprises both domestically and internationally in the field of artificial intelligence have opened up vibrant opportunities for technological progress and practical application of large language models.

Full Text

Preamble

In recent years, large language model technology has advanced by leaps and bounds. The rapid technological development and expansion of internet enterprises both domestically and internationally in the field of artificial intelligence have opened up vibrant opportunities for technological progress and practical application of large language models.

1. Definition and Characteristics of Large Language Models

Large language models (LLMs) are pre-trained language models (PLMs) trained on massive corpora, representing an approach to natural language processing (NLP). Early large language models were primarily built on foundational architectures such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). However, these models could not effectively extract and generate text features. With the introduction of generative models in large language models, the field achieved breakthrough performance in NLP tasks by learning the probability distribution of language to generate new text similar to training data [1].

Large language models generally refer to language models containing hundreds of billions of parameters trained on large-scale corpora. The most representative architecture is the Transformer model, a feedforward fully connected neural network that uses multi-head self-attention mechanisms to construct an encoder-decoder structure [5]. The use of Transformer as a feature extractor endows models with characteristics of easy parallelization, ability to capture long-distance dependencies, and strong comprehensive feature extraction capabilities, along with certain multilingual feature extraction abilities [6].

Large language models follow a processing pipeline of pre-training and adaptive fine-tuning [2]. Pre-training plays a crucial role in LLMs, enabling general knowledge from massive corpora to be embedded into model parameters through language modeling and denoising autoencoding [3]. Adaptive fine-tuning subsequently adjusts the model for specific downstream tasks to achieve better performance and accelerate training.

Based on their massive parameters and architectural features, large language models exhibit emergent capabilities and multilingual extraction abilities in practice. First, emergent capabilities manifest in three main aspects: (1) In-context learning, where models can generate expected outputs for test instances through input text sequences; (2) Instruction-following ability, which improves generalization through instruction fine-tuning; and (3) Multi-step complex task processing capability, which solves complex tasks through chain-of-thought reasoning strategies [4]. Second, multilingual extraction ability is primarily demonstrated in the Transformer architecture, which utilizes mechanisms that enable cross-lingual feature extraction.

2. Technical Trends and Business Layouts of Internet Enterprises at Home and Abroad

Since the proposal of the Transformer neural network and the achievement of state-of-the-art results by the BERT model across multiple NLP domains, pre-trained language models have expanded from monolingual to cross-lingual, multimodal, and lightweight task paradigms [7]. This technological shift into new modalities has also driven a new landscape in the global AI field (see Table 1 for details).

2.1 Foreign Internet Giants

Foreign internet giants have been highly active, pursuing Artificial General Intelligence (AGI) capabilities in technology development while integrating large language models into existing products in their business layouts. Technologically, foreign internet enterprises started early and developed rapidly in large language models. Although they have different technical focuses, all pursue AGI capabilities, striving to achieve human-level intelligence. Google emphasizes technological exploration and multi-path development, launching the BERT model in 2018, T5 in 2019, LaMDA in 2021, and PaLM in 2022. OpenAI focuses on revis-

ing and refining the GPT model, achieving significant progress from GPT-1 in 2018 to the latest GPT-4 in terms of both model scale and adaptability. Meta emphasizes model performance enhancement; its latest LLaMA-13B model can run on a single GPU, offering small scale with high performance.

In business deployment, foreign internet enterprises have broad and diverse business models, mostly relying on their existing commercial ecosystems. OpenAI, primarily engaged in AI research, benefits from natural user data accumulation as one of the earliest companies to open large language models to the public, with business models oriented toward APIs, subscriptions, and technical partnerships. Microsoft's commercial footprint spans productivity, intelligent cloud, and personal computing, integrating GPT-4 into search engine New Bing and Office products like Microsoft 365 while launching Security Copilot. Amazon focuses on e-commerce and cloud computing, expecting to develop proprietary large language models to serve enterprise customers within its existing model. After launching the chatbot Bard, Google introduced the Google Cloud Security AI Workbench based on the Sec-PaLM model to enhance threat detection and analysis.

2.2 Domestic Internet Companies

Domestic internet companies are actively catching up, emphasizing independent R&D in technology development and full-stack implementation in business layout. Technologically, domestic internet companies started later in large language model research but have developed rapidly, with most choosing independent development. Baidu has continuously refined its ERNIE model since its first release in 2019, achieving ERNIE 3.0 for the Wenxin Yiyan application—a general-purpose model across different task paradigms and domains. Alibaba has released Tongyi M6, Tongyi-AliceMind, Tongyi vision models, and Tongyi Qianwen, exploring its own technical path. Huawei released the PanGu-Σ model based on the MindSpore framework, featuring large parameter scale and a Transformer decoder architecture expanded through Random Routed Experts (RRE), providing new ideas for domain model evolution. Tencent has built its Hunyuan large model based on the Taiji platform.

In business deployment, domestic enterprises focus on R&D and full-stack implementation, from independent model construction to application scenario planning mostly relying on their existing internet ecosystems, though current applications remain primarily at the chatbot stage. Baidu released Wenxin Yiyan and plans to provide services through Baidu Cloud, with a business architecture comprising Baidu Brain + Platform + Applications. 360 Zhinao will be deeply integrated with browsers, digital assistants, Soda Office, and intelligent marketing scenarios. Huawei's PanGu-Σ offers unique performance advantages in distributed clusters and Huawei's full-stack environment, enabling fine-tuning on application data for open-domain dialogue, question answering, machine translation, and code generation. Alibaba's Tongyi Qianwen has accumulated full-stack "AI + Cloud Computing" technical strength from the Feitian cloud operating

system to self-developed chips and intelligent computing platforms.

2.3 Future Prospects

Large language models are undergoing unprecedented transformative development in both technological advancement and commercial application, with prospects encompassing both technical breakthroughs and innovative commercial scenarios.

First, in terms of technical prospects, LLMs need to overcome computational and model limitations to enhance intelligence levels. Currently dominated by generative models, LLMs undergo pre-training and fine-tuning through massive data parameters. Despite vigorous development, large-scale data processing depends heavily on computational resources. The processing capabilities of computers, GPUs, and TPUs both support and limit continuous model scaling. How to improve computing power and refine models has become a key challenge for LLMs and NLP.

Second, in commercial prospects, LLMs need broad deployment across all application endpoints to improve content production efficiency. Current commercial applications focus on existing internet layouts, but LLMs' natural language processing capabilities, large-scale intelligent databases, and generation abilities for text and images endow them with production capacity. From a production perspective, LLM applications in AI enable large-scale rapid generation of text and image content, making visions like the metaverse and digital twins possible. From a consumption perspective, LLMs' natural language processing capabilities can provide tiered and customized products for governments, enterprises, and individuals, achieving business model expansion and innovation.

Model	Parameter Scale	Company	Application Features
Wenxin Yiyan	260 billion	Baidu	Services provided via Baidu Cloud; architecture includes Baidu Brain + Platform + Applications
360 Zhinao	-	360	Deep integration with browsers, digital assistants, Soda Office, intelligent marketing
PanGu-Σ	1.085 trillion	Huawei	Unique performance in distributed clusters and Huawei full-stack; fine-tunable for open-domain dialogue, Q&A, machine translation, code generation

Model	Parameter Scale	Company	Application Features
Tongyi Qianwen	1.2 trillion	Alibaba	Full-stack "AI + Cloud Computing" strength from Feitian OS to self-developed chips and intelligent computing platforms
HunYuan-NLP 1T	-	Tencent	Layout across application and model layers, supporting more scenarios
-	11 billion	JD.com	Focuses on task-oriented intelligent dialogue and interaction via voice, text, digital humans for complex tasks, empowering industrial AI
ChatGPT	100 trillion	OpenAI	Supports arbitrary-length multimodal input/output, multilingual interaction, embeddable in other applications
-	-	Microsoft	Copilot for Word, PowerPoint, Excel, Outlook; New Bing search engine with multimodal interaction
Bard	137 billion	Google	Provides latest, high-quality answers using LLMs and web information
Sec-PaLM	-	Google	Processes proprietary threat intelligence from Google and Mandiant to help identify/contain malicious activities and coordinate incident response
LLaMA	7B, 13B, 33B, 65B	Meta	Small size, high performance, runnable on single GPU

References

[1][3][4] Zhao W X, Zhou K, Li J, et al. A Survey of Large Language Models[J]. arXiv preprint arXiv, 2023.
[2][5] Yue Z Y, Ye X, Liu R H. Research on Pre-trained Language Model Techniques[J]. Journal of Chinese Information Processing, 2021(9): 15-29.
[6] Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT?[J]. arXiv preprint arXiv, 2019.

[7] Abudukelimu·Abulizi, Zhang Y N, Alimujiang·Yasen, et al. Research on Extended Models of Pre-trained Language Models[J]. Computer Science, 2022(S2): 43-54.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.