Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 9, No. 2, 2024

Security Challenges and Reflections on Large Models

Xuan Huang [1, 2], Linyi Huang [1, 2], Guowei Tong [1, 2], Xundao Zhou [1, 2] and Jianheng Luo [1, 2, *]

[1] CEPREI, Guangzhou, Guangdong, China
[2] Key Laboratory of MIIT for Intelligent Products Testing and Reliability, Guangzhou, Guangdong, China

*** Corresponding author:** Jianheng Luo

Abstract: The rapid development of large-scale AI models has revolutionized the technology industry, offering unprecedented opportunities for innovation across various sectors. This paper discusses the emergence of the "Hundred Model War" and the significant growth in large models, highlighting the potential for transformative applications in vertical fields such as automotive, medical, and finance. However, we also identify significant challenges, including safety ethics, governance systems, and the vulnerability of models to malicious attacks. The paper concludes with a call for the establishment of a comprehensive ethical governance system, improved safety supervision mechanisms, and the development of a public technology resource support platform to ensure the sustainable and healthy development of AI technologies.

Keywords: Artificial Intelligence; Large Models; Security Challenges; Ethical Governance.

1. Introduction

On May 14, 2024, OpenAI released its latest AI model GPT-4o [1], which has capabilities spanning speech, text, and vision, once again causing a sensation in the technology industry. In fact, since the official release of the AI chatbot ChatGPT in 2022, news of the landing of AI models by major domestic and foreign enterprises such as Google, Microsoft, Meta, Anthropic, Baidu, Alibaba, Huawei, and iFlytek has emerged one after another, ushering in the outbreak of the "Hundred Model War" in the industry. The general artificial intelligence large model and its vertical field applications are regarded as a new development trend in the artificial intelligence industry.

2. Large Models Lead New Opportunities for Ai Development

ChatGPT ignites the internet, leading to explosive growth of domestic large models. According to data from the National Data Administration in March 2024, there are over 100 large models in China with a parameter scale of over 1 billion. According to incomplete statistics from the Daily Economic News, as of the end of April 2024, a total of 305 large models have been launched in China, of which about 140 have completed the registration of generative artificial intelligence services. Geographically speaking, it is mainly concentrated in the locations of national artificial intelligence innovation and application pilot zones such as Beijing, Shanghai, Guangzhou, Shenzhen, and Hangzhou. In terms of technology, the focus is mainly on natural language processing and multimodal fields, while the number of computer vision and intelligent speech models is relatively small.

General large-scale models are competing fiercely, with various applications in full swing. The universal large model is the main battlefield of the "Hundred Model Battle", with major technology companies, universities, research institutions, and others laying out universal large models. First, in terms of enterprises, Baidu, Ali, the Dark Side of the Moon, iFLYTEK, Baichuan, Step Star, etc. launched ChatGPT like C-end applications based on ERNIE Bot[2], Tongyi Qianwen, Kimi, iFLYTEK Spark, Baichuan, and Yuewen, respectively, while Huawei, Tencent, Shangtang, etc. launched Pangu, Hunyuan, Ririxin, and other large models for internal and B-end applications; Secondly, in terms of universities, Tsinghua University and Fudan University have respectively launched open-source models such as ChatGLM3 and MOSS; Third, in terms of research institutions, Shanghai Artificial Intelligence Laboratory, IDEA Research Institute, Pengcheng Laboratory, and the Institute of Automation of the Chinese Academy of Sciences launched universal models such as InternLM[3], MindBot, Pengcheng Brain, and Zidong Taichu.

Vertical field applications are flourishing, and large models are expected to empower thousands of industries. Compared to the general large model market abroad, the development opportunity of domestic large models lies in vertical field applications. Currently, there are large models released in multiple fields such as automotive, medical, manufacturing, and finance, mostly based on open-source large models for vertical domain fine-tuning. In the field of automobiles and transportation, multiple car companies have successively launched large-scale model car competitions. Ideal "Mind GPT", NIO "NOMIGPT", Geely Xingrui AI large model, GAC AI large model platform, SenseTime Jueying, etc., mainly targeting driving scenarios such as in car intelligent voice interaction, intelligent driving, intelligent networking, and intelligent cockpit; Huawei Pangu Automotive Model, mainly targeting business scenarios such as automotive design, production, marketing, and research and development; Jiadu Zhixing Transportation mainly focuses on intelligent operation and management scenarios for road and rail transportation. In the fields of medicine and biology, there have emerged large models for medical consultation and psychological counseling, such as "Baidu Lingyi", Harbin Institute of Technology "Bencao", South China University of Technology "Bianque" and "Lingxin", Hong Kong University of Science and Technology Shenzhen "Huatuo", as well as large models for drug design, biomolecular computing, genomics computing, and other applications such as Pangu drug molecular model and Wenxin Biocomputing.

3. Challenges and Hidden Dangers in The Development of Large-Scale Models

Concerns about safety ethics have arisen, and the governance system still needs to be established. The current defense capabilities of large models are generally insufficient and vulnerable to malicious attacks such as command attacks, prompt injection, and backdoor attacks, which may produce outputs that do not conform to human values, including discriminatory speech, insults, and content that violates laws and ethics. For example, on the day of the birth of GPT-4o, the jailbreak attack paradigm targeting GPT-4o was made public, which could allow GPT-4o to freely leak dangerous information, including sensitive topics such as the production of dangerous goods and prohibited drugs. In response to the above risks, countries around the world have begun to formulate various inspection and regulatory policies for large models. In April 2023, the United States released the "Solicitation of Opinions on Artificial Intelligence Accountability Policies" to solicit opinions on security risk assessment, accountability, and other related measures for artificial intelligence systems such as ChatGPT. In March 2024, the European Parliament voted to pass the Artificial Intelligence Act, which proposes a regulatory model based on risk levels and has a dedicated chapter to regulate general artificial intelligence models. Domestically, in July 2023, the Cyberspace Administration of China and seven other departments jointly released the "Interim Measures for the Management of Generative Artificial Intelligence Services", proposing a top-level design concept for the security and compliance governance system and regulatory measures of generative artificial intelligence technology services, in order to promote the healthy development and standardized application of generative artificial intelligence. In May 2024, the National Cybersecurity Standards Committee released the "Basic Requirements for Network Security Technology Generative Artificial Intelligence Service Security" (draft for comments), aiming to help service providers clarify the network security baseline of generative artificial intelligence services and improve service security levels. In addition, in terms of research institutions, institutions such as Tsinghua University and Beijing Jiaotong University have successively launched training and evaluation datasets for aligning and governing Chinese values, such as Safety-Prompts and Cvalues[4]. The Fifth Institute of the Ministry of Industry and Information Technology has designed five ethical and moral evaluation datasets based on Chinese values and four types of security attack samples based on the PromptBench benchmark. However, overall, there is currently no comprehensive plan and ethical governance mechanism for ensuring the safety and compliance of large-scale models. The industry mainly relies on self declaration, application filing, complaint reporting, and post event disposal. The pre supervision and guarantee work for the safety and compliance of large-scale models and service quality still needs to be carried out.

Ranking competition dominates, and evaluation benchmarks still need to be improved. Large scale model evaluation is an important part of understanding, recognizing, and evaluating its performance level. Currently, multiple institutions at home and abroad have proposed various types of evaluation benchmarks, including MMLU[5] (Multidisciplinary English), C-Eval[6] (Multidisciplinary Chinese), SQuAD[7] (English text), SuperCLUE[8] (Comprehensive Chinese), etc., and have put forward multiple evaluation indicators such as F1, BELU, ROUGE, etc. The industry relies on various benchmarks to evaluate and publish models, showing a trend of fast updates to rankings and increasing scores in indicators. However, in terms of practical application experience, there is still a significant gap between the performance of large models and human expectations, which reflects the imperfect evaluation benchmarks and ranking mechanisms in existence. One is excessive reliance on benchmark testing, which simplifies complex and comprehensive evaluation requirements into a single number for a specific indicator, and the industry has not yet established a complete testing and evaluation system; Secondly, after the evaluation and ranking of the large model, it can be fine tuned and optimized based on open-source benchmarks to increase the evaluation score; The third issue is that the evaluation rankings are not exhaustive and the evaluation criteria are inconsistent, resulting in significant differences in the results of large models with similar overall abilities on different rankings. The industry urgently needs scientific, fair, open, and standardized evaluation methods and tools.

Key chips are stuck and there is a shortage of high-quality Chinese datasets. The success of large models cannot be achieved without high-performance computing power and high-quality data. With the continuous popularity of large models and generative AI, the demand for high-performance chips in the market has skyrocketed. Nvidia A100 and H100 chips, which can meet the dual precision floating-point computing capability, are in a monopoly position in the market and are subject to export controls by the US government. Therefore, high-performance chips have always been in short supply. Currently, domestic GPU chips mainly support single precision floating-point computing, and there is a certain gap in core performance indicators. At the same time, due to issues such as the chip programming ecosystem and system level computing platforms, the support of large model training software tools and deep learning frameworks for domestic chips is extremely limited, resulting in a large amount of operator adaptation work for domestic computing power, greatly restricting the research and application of large models in China. In the future, leading enterprises need to take the lead in collaborating with industry and academia to carry out full stack adaptation and migration. In addition, compared to English corpora, there is a severe lack of high-quality training corpora available in the existing Chinese community, making it difficult to fully train and align Chinese values for domestically produced large-scale models with extremely large parameters.

4. Thoughts and Suggestions

Build a large-scale ethical governance system and improve safety supervision mechanisms. It is suggested to research and develop a governance and regulatory mechanism for large models, focusing on the standardization of training data, product launch evaluation and certification, content generation supervision, information privacy protection, rights and security protection, application scenario creation, service quality evaluation, and management platform construction. A large model and generative artificial intelligence technology ethical governance system and security supervision strategy should be constructed to encourage innovative application of large model technology while complying with public order,

good customs, and social ethics, ensuring the sustainable and healthy development of new artificial intelligence technologies.

Improve the safety and compliance evaluation system to safeguard the development of the industry. Unlike the brushing ecology constructed by academia and industry, industry organizations and third-party institutions should develop a set of security compliance testing and evaluation standards and evaluation datasets that cover ethics, value alignment, legal compliance, security protection, privacy protection, authenticity and reliability, actively participate in the formulation of standards related to large models, and contribute to the development of the large model industry.

Based on vertical field applications, build a public technology resource support platform. The key point of the international competitive strategy in the Hundred Model Battle lies in the number of native AI applications that large models possess, rather than the number of large models. Vertical domain applications will be the core of China's large-scale model ecosystem. Build a training and evaluation public service platform for vertical domain applications of large models. Encourage research institutes and enterprises to increase their R&D investment in domestically produced high-performance chips and reduce their dependence on imported chips through special funds, subsidy policies, and other forms. Develop multi-objective joint reinforcement learning human feedback evaluation criteria for large vertical domain models. Carry out small-scale pilot projects for vertical large models in fields such as automotive, manufacturing, and healthcare, explore sandbox regulatory mechanisms, and provide public resource technical support and testing evaluation services such as computing power, algorithms, and data to enhance the accuracy, professionalism, safety, and credibility of large models and their applications in vertical fields.

Acknowledgments

This paper was supported by 2022 Industrial Technology Basic Public Service Platform Project of China (No.2022-228-219). Luo Jianheng is the corresponding author.

References

[1] OpenAI. (2024). GPT-4o: A Multimodal AI Model. [Online] Available: https://openai.com/2024/gpt-4o [Accessed: July 2024].

[2] Zhang Bolin; Tu Zhiying; Hang Shaoshi, Chu Dianhui, Xu Xiaofei, “Conco-ERNIE: Complex [2] User Intent Detect Model for Smart Healthcare Cognitive Bot”, ACM Transactions on Internet Technology, Vol.1, 2023.

[3] Pan Zhang, Xiaoyi Dong, Yuhang Zang, et al. “InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output”, arXiv: 2407. 03320.

[4] Guohai Xu, Jiayi Liu, Ming Yan, et al. “CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility”, arXiv:2307.09705.

[5] H Dan, C Burns, S Basart, et al. “Measuring Massive Multitask Language Understanding”, Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), 2021.

[6] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, et al. “C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models”, arXiv:2305.08322.

[7] Pranav Rajpurkar, Robin Jia, Percy Liang, “Know What You Don't Know: Unanswerable Questions for SquAD”, arXiv: 1806.03822.

[8] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich, “SuperGlue: Learning Feature Matching with Graph Neural Networks”, arXiv:1911.11763.