Journal of Artificial Intelligence and Technology, 2023, 3, 95-99
https://doi.org/10.37965/jait.2023.0298

ISTP
RESEARCH ARTICLE

Research on Miniaturization Trend of ChatGPT Technology Model

Xuemei Shi,[1,2] Xiaoguang Du,[3] and Xiedong Song[4]
[1]Computer College of Huainan Normal University, Huainan, China
[2]College of Computing and Information Technologies, National University, Philippines, Philippines
[3]Biological Engineering of Huainan Normal University, Huainan, China
[4]Department of Electronic Information, Yantai City Vocational College of Science and Technology, Yantai, China

(Received 01 May 2023; Revised 12 May 2023; Accepted 12 May 2023; Published online 17 May 2023)

Abstract: Miniaturization and micro-miniaturization are trends in technology models, such as ChatGPT. These trends have the potential to enhance the practicality and professionalism of the model, as well as making them more widely accessible. Consequently, more individuals and organizations can leverage these technologies, and their impacts can be significant. Notably, miniaturization and micro-miniaturization can decrease the size of the model and the computing resources required, thus resulting the widespread use and development of artificial intelligence technology. Moreover, they can boost the speed of model operation and training efficiency, thereby improving the practicality and efficacy of applications. Ultimately, this trend will have a profound impact on diverse fields, including scientific research, education, coaching, medical care, and daily life.

Keywords: class ChatGPT model; miniaturization; practicality

I. INTRODUCTION

Miniaturization and micro-miniaturization are trends in technology models, such as ChatGPT. These trends the potential to enhance the practicality and professionalism of the model, as well as making them more widely accessible. This discusses the rise of the ChatGPT. Several Chatbot models are commonly used.

A. THE RISE OF THE ChatGPT-LIKE MODEL

The increasing popularity of chatbot models in recent years is largely attributed to the advancements in using neural networks for natural language processing (NLP) [1]. Among the most renowned models is the generative pre-trained transformer (GPT) model developed by OpenAI, which is trained in an unsupervised manner by learning the language structure and rules from extensive corpora and applying this knowledge to produce new text. Built on the Transformer architecture, the GPT model is a neural network model capable of generating high-quality natural language text [2].

Due to their ability to engage in natural conversation with humans, chatbots have numerous applications [3–5], such as online customer service, virtual assistants, and speech recognition systems. They have enhanced user experience and satisfaction.

The ChatGPT-like models are technologies that have emerged to meet the folloing demands:

Deep learning technology: The advancement of deep learning technology [6] has provided powerful techniques for optimizing models of ChatGPT-like technologies. These neural network-based models can be optimized by deep learning techniques that continuously refine the model structure and parameters to enhance the model's accuracy and performance.

NLP technology: The advancement of NLP technology has significantly contributed to the development of ChatGPT-like models. NLP techniques enable the processing and analysis of text data, extracting valuable semantic and emotional information that provides more accurate guidance and optimization for text generated by ChatGPT-like models.

Large-scale corpora: The advent of the internet and the accumulation of data have resulted in the collection and sharing of an increasing amount of textual data. This data serves as a foundation for the development of data-driven NLP technology. The training of ChatGPT-like models requires an extensive volume of textual data, and the exponential growth of textual data on the internet provides abundant sources for the training of such models.

In conclusion, the emergence of ChatGPT-like models [7,8] is the outcome of the synergistic action of various factors and an inevitable byproduct of the development of NLP and machine learning technology [9]. As technology continues to advance and its application is increasingly promoted, the development prospects of ChatGPT-like models will undoubtedly become even more extensive.

B. SEVERAL COMMONLY USED CHATGPT-LIKE MODELS

ChatGPT, as an AI assistant, is capable of performing a variety of NLP tasks, including question answering, conversation, and text generation. The following are several GPT-based language models that are frequently used:

GPT-1: The GPT-1 model, developed by OpenAI, was the first of its kind, boasting 150 million parameters, and has been utilized for generating a diverse range of text, such as news articles, short stories, and poetry.

GPT-2: The GPT-2, OpenAI's second GPT model, has a parameter range of 150 million to 1.5 billion and is currently

Corresponding author: Xuemei Shi (e-mail: 784909013@qq.com).

one of the most popular language models. It demonstrates excellent performance in various NLP tasks, particularly in generating natural and fluent language.

GPT-3: GPT-3, with its 1.75 trillion parameters, is OpenAI's third and largest GPT model to date, suitable for a wide range of NLP tasks such as machine translation, conversation generation, and question-answering systems.

DialoGPT: A GPT-based language model is designed for generating conversations and fine-tuned based on GPT-2. DialoGPT excels at producing coherent conversations and finds wide applications in intelligent customer service and chatbots.

MiniLM: MiniLM is a GPT model with a lightweight architecture released by Microsoft, comprising only 60,000 parameters. Despite its relatively small size, MiniLM offers high-quality generation while requiring low computational and storage costs. As such, it is particularly suited to resource-limited scenarios.

This study has contributed in the following ways:

Miniaturization and micronization are the only way for practical applications, and the study will make an important contribution to the field of artificial intelligence technology, looking into the future.

The rest of the paper is organized as follows. Section II presents the related work. Section III discusses the extensively used ChatGPT-like models. Section IV introduces the maturity of the ChatGPT-like models. Section V presents conclusions and discussions.

II. RELATED WORK

ChatGPT [10] is an AI language model that is based on the GPT-3.5 architecture. Its related work primarily focuses on the following areas:

OpenAI GPT series models: In 2018, OpenAI released the GPT model and later introduced versions such as GPT-2 and GPT-3. These models possess strong capabilities for generating and understanding text and are widely utilized in fields including chatbots and NLP.

Bidirectional encoder representations from transformers (BERT) model: BERT, a Transformer-based language model developed by Google, can perform tasks such as text classification, question answering, and language reasoning. Its emergence has had a significant impact on the field of NLP and has also provided valuable guidance for the development of ChatGPT.

GShard model: GShard is Google's newest distributed training platform for large-scale Transformers, which can train models with billions of parameters simultaneously. This technology can improve the performance and efficiency of ChatGPT, making it applicable to a wider range of scenarios.

Chinese GPT model: Chinese GPT is a language model for the Chinese language, introduced by Huawei Cloud. While its base model is similar to the English GPT-2, it has been specifically optimized for the unique characteristics of the Chinese language. The emergence of the Chinese GPT has provided valuable support and guidance for NLP in Chinese.

Joint Laboratory of HFL and iFLYTEK Language Model: This laboratory has achieved significant breakthroughs in the field of language models, introducing models based on Transformer and long short-term memory (LSTM) is a variant of recurrent neural networks (RNN) used for handling sequential data. Architectures, which have found extensive applications in NLP, machine translation, text classification, and other domains.

PaddleNLP: PaddleNLP is a NLP toolkit launched by Baidu PaddlePaddle. It includes Transformer-based models, BERT models, and Enhanced Representation through Knowledge Integration (ERNIE) models, as well as various pretraining models and training tools, making the development of ChatGPT more convenient and efficient.

In conclusion, all countries research and development in language models and NLP provide valuable references and support for the development of ChatGPT. These works offer strong support for ChatGPT to play a greater role in application scenarios.

III. EXTENSIVE USE OF CHATGPT-LIKE MODELS

ChatGPT-like models have been widely utilized in the field of NLP and related domains. They have brought numerous conveniences and innovative solutions to people's lives and work.

A. TREND CHARCTERISTICES OF MINIATURIZATION AND MINIATURIZATION

The miniaturization and micronization trend of ChatGPT-like models can be identified by several key features:

Firstly, incremental learning has emerged as an alternative to traditional model training. This method enables continuous expansion and updating of the model's knowledge while maintaining its performance. As a result, it significantly reduces the computational and storage costs of the model, making it suitable for resource-constrained scenarios.

Secondly, model pruning and quantization are effective techniques for reducing the size of large models. Model pruning can remove redundant weights and layers, thereby decreasing the number of model parameters. Additionally, model quantization can convert model parameters into low-bit width data types, further reducing the model's size and computational cost.

Thirdly, the distillation technique is another effective approach for model miniaturization. This technique involves transferring the knowledge of a large model to a smaller one, while still retaining the original model's performance. As a result, it reduces the computational cost and size of the model.

Fourthly, decentralized computing frameworks are gaining popularity as a solution to the issue of computing resources for large models. These frameworks enable the utilization of edge devices' computing resources, reducing the time and cost of model training and inference.

Finally, multitask learning has been shown to be effective in improving the efficiency and generalization ability of ChatGPT-like models. By leveraging the correlation and shared knowledge between different NLP tasks, this approach reduces the model's size and computational cost while maintaining performance across various tasks.

B. PRACTICALITY BROUGHT BY MINIATURIZED AND MINIATURIZED CHATGPT

The miniaturization and micronization of ChatGPT models bring many practical benefits, including:

Table I. The miniaturized GPT-2 model (124 M parameters) compared with the large model (345 M parameters)

GPT-2 model	The classification accuracy of the GPT-2 model	GPT-2 model inference speed	The bilingual evaluation understudy (BLEU) score of the GPT-2 model
Reduced by three times [14]	Reduced by 1.6% [15]	It is more than three times faster now [16]	Less than one point lower [17]

Saving computing resources: Miniaturized and micronized ChatGPT models can significantly reduce model size and computational cost, making them able to run on resource-constrained devices such as mobile and embedded devices.

Improving running speed: Due to the reduced computational cost of miniaturized and micronized ChatGPT models, they can complete inference [11] and training tasks [12] faster.

Reducing costs: Using miniaturized and micronized ChatGPT models can reduce hardware and software costs, which is crucial for organizations and enterprises that need to deploy NLP applications in resource-constrained environments.

Improving deployment flexibility: Miniaturized and micronized ChatGPT models can be more easily deployed in a variety of devices and applications, thus improving the deployment flexibility [13] of the models.

The practical statistics of ChatGPT are shown in Table I.

In summary, the miniaturization and micronization of ChatGPT models have rendered NLP technology more practical by enabling its application across a broader range of devices and scenarios without compromising on performance and efficiency.

C. PROFESSIONALISM BROUGHT BY MINIATURIZED AND MINIATURIZED CHATGPT

The miniaturization and micronization of ChatGPT technology has made this NLP model easier to use in various application areas, resulting in increased professionalism [18]. Specifically, reducing the model size and computational cost to an acceptable level enables miniaturized and micronized ChatGPT models to be better applied in resource-limited scenarios, such as mobile devices, IoT devices, and edge computing devices. These scenarios often require models with smaller size and computational capability to meet device constraints. Moreover, miniaturized and micronized ChatGPT models can be better applied to small-scale enterprises and individual developers, enabling them to leverage this powerful NLP technology to build their own applications. Therefore, miniaturized and micronized ChatGPT models have brought broader applications to the professional field of NLP, improving its professionalism [19].

D. POPULARITY BROUGHT BY MINIATURIZED AND MINIATURIZED CHATGPT

The miniaturization and micronization of ChatGPT have facilitated its widespread use by making NLP technology more accessible across diverse scenarios. Traditional large-scale models often demand considerable computational resources and entail high costs, thus limiting the reach and scope of these technologies. However, the miniaturization and micronization techniques have significantly decreased the size and computational complexity of models, rendering them easier to apply on a range of devices, including mobile devices, smart speakers, and smart homes. Furthermore, these technologies provide developers with ample options to customize their models based on their specific requirements and budgets, thereby achieving better balance between performance and cost. These improvements have made ChatGPT technology more popular, accelerating their development and use in practical scenarios.

Table II. Parameter statistics

DistilGPT-2 model	Small GPT-3 model	The original model
2.58 million	460 million	175 billion

The miniaturization and micronization of ChatGPT models can lead to greater accessibility, primarily manifested in the following ways:

Expanded application scenarios: Miniaturized and microized models, capable of running on edge devices, can broaden the application scenarios of ChatGPT models, such as smartphones, tablets, and smart speakers. Since these devices have a high adoption rate, people can conveniently reap the benefits of ChatGPT models.

Reduced barriers and costs: Miniaturized and micronized models can efficiently perform NLP using fewer computational resources and less training data, thus lowering the barriers and costs of using and developing ChatGPT models. This feature is highly advantageous for small- and medium-sized enterprises and individual developers.

Enhanced user experience: Miniaturized and micronized models can execute inferences and respond more quickly, resulting in improved real-time performance and user experience of ChatGPT models. For instance, using miniaturized and micronized ChatGPT models in smart speakers can enable faster response to user voice commands, thereby providing a superior user experience.

The parameter statistics are shown in Table II.

Thus, the miniaturization and micronization of ChatGPT models can increase the accessibility of ChatGPT models, allowing more individuals to benefit from the convenience and advantages that they offer.

IV. A SIGN OF THE MATURITY OF THE CHATGPT-LIKE MODEL

The maturity of ChatGPT-like models is reflected in their successful application across various real-world scenarios and their capacity to generate high-quality responses that are indistinguishable from human-generated ones.

A. MINIATURIZATION AND MICRONIZATION ARE SIGNS OF THE MATURITY OF CHATGPT-LIKE MODELS

Miniaturization and micronization represent crucial stages in the development of ChatGPT models and are among the indications of their maturity. In their early developmental phases, ChatGPT models often require substantial computational resources and data to attain optimal performance, which restricts their practical use. Nonetheless, researchers have begun to explore ways to optimize model structures and training methods to reduce the model size and computational expenses, thereby enhancing the models' practicality and popularity.

Miniaturization and micronization technologies have made it possible to employ ChatGPT models in resource-constrained

environments. For instance, miniaturized and micronized ChatGPT models can operate on edge devices such as smartphones, enabling them to perform NLP functions on the device without requiring data to be transmitted to the cloud for processing. Consequently, data security and processing efficiency are improved. Moreover, miniaturization and micronization technology has enabled ChatGPT models to be utilized in a more extensive range of contexts, including smart homes, intelligent customer service, and intelligent voice assistants.

Thus, it can be asserted that miniaturized and micronized ChatGPT models represent one of the signs of ChatGPT model maturity. Their emergence has expanded the potential application scenarios for ChatGPT models and introduced new prospects and challenges for their development.

B. MINIATURIZATION AND MINIATURIZATION ARE SIGNS OF THE POPULARITY OF CHATGPT-LIKE MODELS

The miniaturization and micronization of ChatGPT models are clear indicators of the increased prevalence of ChatGPT models, as they facilitate their deployment across a wider range of devices and scenarios. Conventional ChatGPT models are typically characterized by hundreds of millions or billions of parameters, necessitating significant computational resources and storage space. As a result, their applicability is limited to devices with abundant resources and high computing power, precluding their use on low-end devices such as mobile devices and embedded systems. However, miniaturization and micronization techniques can effectively reduce model size and computational complexity, thus enabling the deployment and use of ChatGPT models on such devices. This increased accessibility expands the practical value of ChatGPT models across diverse application scenarios, including intelligent customer service, intelligent voice assistants, automatic question answering systems, and others.

C. MINIATURIZATION AND MICRONIZATION ARE THE ONLY WAY FOR THE PRACTICAL APPLICATION OF CHATGPT

Miniaturization and micronization are crucial for enhancing the practicality of ChatGPT models. As these models continue to evolve, their size and number of parameters have been increasing exponentially, leading to high training and deployment costs and limited scope and effectiveness in practical applications.

However, the advent of miniaturization and micronization technology has effectively resolved the issue of model size and computational cost, making ChatGPT models more practical. Techniques such as decentralized computing, model pruning and quantization, incremental learning, multitask learning, and distillation can reduce the size and computational complexity of ChatGPT models to an acceptable level while maintaining high performance and accuracy. This not only benefits large-scale applications of ChatGPT models but also allows their deployment and use on resource-limited edge devices. Therefore, miniaturization and micronization technology are essential to make ChatGPT models practical.

V. CONCLUSION AND DISCUSSION

The trend of miniaturization and micro-miniaturization of ChatGPT-like models has had a significant impact on their efficiency and practicality, making them more accessible and applicable in various real-world scenarios.

A. PROSPECTS FOR THE WIDESPREAD USE OF CHATGPT

The advancement of ChatGPT technology has resulted in significant transformations and innovations [18], with promising and broad application prospects in the field of NLP. The following presents a comprehensive outlook on the widespread use of ChatGPT models:

Voice recognition and intelligent conversation: Class ChatGPT technology has the potential to enhance the accuracy of voice recognition and improve the quality of intelligent conversation. In the future, individuals will be able to interact intelligently with smart homes, intelligent robots, and other devices through voice commands, achieving more advanced human–machine interactions.

Natural language generation and translation: Class ChatGPT technology can produce natural, smooth text, which would result in machine-generated articles, news, novels, and other content of higher quality and readability. Moreover, class ChatGPT technology can help translation systems understand language more accurately, thus boosting the quality and efficiency of translations.

Automatic summarization and text classification: Class ChatGPT technology can automatically summarize and classify text, facilitating rapid understanding of the themes and key points of large volumes of text information. This feature would have crucial practical implications in domains such as news media, financial analysis, and medical diagnosis.

Intelligent customer service and support: Class ChatGPT technology can enable companies to set up intelligent customer service systems that automatically reply to customer inquiries and resolve issues, leading to improved customer satisfaction and organizational efficiency.

Social media analysis and public opinion monitoring: Class ChatGPT technology can automatically analyze and monitor massive amounts of social media data, allowing governments, businesses, and other entities to better understand public attitudes and needs, and make more informed decisions.

In summary, the extensive application of class ChatGPT technology has the potential to bring significant convenience and benefits, and it contributes positively to the intelligentization process of humanity and social progress.

B. CONCLUSIONS AND DISCUSSIONS

The usage of the ChatGPT model is limited in certain scenarios due to its large-scale parameters. Consequently, an increasing number of researchers and applications are exploring ways to downsize the ChatGPT model to adapt to low-power and low-storage capacity devices and environments. The impact of the miniaturization and micro-miniaturization trends of ChatGPT-like technology models can be observed in several aspects:

Firstly, a model with faster inference speed is required in some scenarios with high real-time responsiveness, such as smart homes and IoT devices. Miniaturization and micro-miniaturization can decrease the number of model parameters, resulting in an improved inference speed.

Secondly, for devices with limited memory or storage resources, such as mobile devices and embedded systems, a model with smaller storage space is necessary. Miniaturization and

micro-miniaturization can reduce the number of model parameters, resulting in a smaller storage space.

Expanding application scenarios: Miniaturization and micro-miniaturization can expand the adaptability of the ChatGPT model to a broader range of application scenarios such as voice assistants, robots, automatic translation, emotion analysis, and more.

Reduced cost: The process of miniaturization and micro-miniaturization can decrease the computational and storage costs associated with the model, ultimately rendering the ChatGPT model more accessible and cost-effective.

The effectiveness of the model could be influenced: Although downsizing and miniaturization can decrease the number of model parameters, they may also result in a reduction of model performance. Thus, it is essential to carefully balance the model's size and effectiveness during the downsizing and miniaturization process.

To summarize, the downsizing and miniaturization of ChatGPT-like models are currently hot research topics in the field of NLP [19]. This approach can enhance the adaptability of ChatGPT models to practical application scenarios and broaden their commercial prospects.

Furthermore, downsizing and miniaturization represent significant trends in the development of AI technology and will play crucial roles in future applications [20–22].

References

[1] K. Roose, "The brilliance and weirdness of ChatGPT," The New York Times. Accessed on Dec. 9, 2022.

[2] S. Lock, "What is AI chatbot phenomenon ChatGPT and could it replace humans?," The Guardian. 2022.

[3] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," IEEE Trans. Autom. Sci. Eng., vol. 12, no. 2, pp. 398–409, 2015.

[4] T. H. Kung and M. Cheatham, "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models," PLOS Digit. Health, vol. 2023, no. 3. DOI: 10.1371/journal.pdig.0000198. Accessed on Feb. 9, 2023.

[5] B. D. Lund and T. Wang, "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?," Library Hi Tech News, Vol. ahead-of-print No. ahead-of-print. DOI: 10.1108/LHTN-01-2023-0009.

[6] J. Loeffler, "Personalized, learning: Artificial intelligence and education in the future," 2018. Available: https://interestingengineering.com/personalized-learning-artifi-cial-intelligence-and-education-in-the-fututure.

[7] A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM J. Res. Develop., vol. 3, no. 3, pp. 210–229, 1959.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proc. Mach. Learn. Res., vol. 81, pp. 1–15, 2018.

[9] I. A. P. Wogu, S. Misra, P. A. Assibong, E. F. Olu-Owolabi, R. Maskeli□nas, and R. Damasevicius, "Artificial intelligence, smart classrooms and online education in the 21st century: implications for human development," J. Cases Inf. Technol., vol. 21, no. 3, 2019.

[10] M. A. AlAfnan, S. Dishari, M. Jovic, and K. Lomidze, "ChatGPT as an educational tool: opportunities, challenges, and recommendations for communication, business writing, and composition courses," J. Artif. Intell. Technol., vol. 3, pp. 60–68, 2023.

[11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: a lite BERT for self-supervised learning of language representations," *arXiv:1909.11942*, 2019.

[12] D. Adiwardana, M. T. Luong, Q. Liang, Q. V. Le, and O. Vinyals, "Towards a human-like open-domain chatbot," arXiv preprint *arXiv:2001.09977*, 2020.

[13] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," *arXiv:1909.10351*, 2020.

[14] I. Najdenkoska, X. Zhen, and M. Worring, "Meta learning to bridge vision and language models for multimodal few-shot learning," Comput. Vis. Pattern Recog., *arXiv:2302.14794*, 2023.

[15] H. H. Thorp, "ChatGPT is fun, but not an author," Science, vol. 379, no. 6630, Jan. 2023.

[16] E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: five priorities for research," Nature, 2023. DOI: 10.1038/d41586-023-00288-7.

[17] V. Kyrylov, and D. Chaplynskyi, "GPT-2 metadata pretraining towards instruction finetuning for ukrainian," In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, Dubrovnik, Croatia. Association for Computational Linguistics, pp. 32–39, 2023.

[18] S. B. Patel and K. Lam, "ChatGPT: the future of discharge summaries?," 2023. DOI: 10.1016/S2589-7500(23)00021-3.

[19] Y. Shen and L. Heacock, "ChatGPT and other large language models are double-edged Swords," 2023. DOI: 10.1148/radiol.230163.

[20] W. Jiao, W. Wang, J.-T. Huang, X. Wang, Z. Tu, Tencent AI Lab, "Is ChatGPT a good translator? Yes with GPT-4 as the engine," arXiv: submit/4797616. Accessed on Mar. 19, 2023.

[21] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," Comput. Sci. Comput. Lang., 2023. DOI: 10.48550/arXiv.2302.04023.

[22] A. Gilson, C. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, "How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment," 2022. https://preprints.jmir.org/preprint/45312.