ISSN:2790-1688 Volume-6-(2023)

Exploring the Opportunities and Challenges of Developing Large AI Models and their Commercialization

Chenzimo Ran

Sendelta international academy, Shenzhen,518108, China

Abstract. This article explores the evolution and impact of large models in artificial intelligence (AI), with a focus on their role in advancing fields like natural language processing and image recognition. It categorizes this evolution into three stages: deep learning revival, big data and distributed computing, and self-supervised learning. The capabilities of large-scale models, especially in generating AI content, are discussed, along with the role of small models in personalized enterprise requirements and resource-limited environments. Discusses the development of large AI models and the opportunities they bring, as well as the importance of considering user needs when transitioning from a model to a product. The commercialization of AI products is also explored, with attention paid to the details of this process. The future of AI is predicted to be a dynamic balance between large and small models, depending on task requirements and resource limitations. The article concludes with a discussion on the commercialization of AI products, emphasizing the importance of data, algorithms, user-centric design, and viable business models.

Keywords: Large AI Models; GPT; Commercial AI.

1. Introduction

Large models refer to deep learning models with massive numbers of parameters, which usually require distributed computing and special hardware accelerators for training and inference. In recent years, with the development of deep learning technology and the increase in computing resources, large models have achieved many important results in natural language processing, image recognition, speech recognition, and other fields, becoming an important driving force for the development of artificial intelligence. This article will elaborate on the theory, development, and achievements of large models in detail. The development history of large models can be divided into three stages: the first stage is the deep learning revival stage starting from 2012, which mainly improves the model's expression ability and prediction performance by increasing the depth and number of layers, such as AlexNet, VGG, GoogLeNet, and other models. The second stage is the big data and distributed computing stage starting from 2015, which mainly improves the model's generalization ability and complexity by increasing the number of model parameters and training data, such as ResNet, Inception, and other models. The third stage is the self-supervised learning stage starting from 2018, which mainly improves the model's expression ability and generalization ability by utilizing large-scale unlabeled data for pre-training, such as BERT, GPT, T5, and other models.

In these three stages, the development of large models has experienced breakthroughs in many key technologies, including convolutional neural networks, residual connections, batch normalization, and distributed training, which have provided important support for the training and inference of large models. At the same time, the development of large models has also promoted the development of computer hardware, such as the emergence of special hardware accelerators such as GPUs and TPUs, greatly improving the training speed and efficiency of large models.

2. Large Models

2.1 Theoretical foundation of large models

The theoretical foundation of large models is deep learning, which is a machine learning method based on neural networks. It extracts high-order features through multiple layers of nonlinear

transformations to model and predict complex tasks. The core idea of deep learning is to train model parameters by optimizing the loss function, so that the model can better fit the training data and have good generalization ability. Deep learning models usually consist of multiple layers, each layer including a group of neurons and a group of weight parameters, which are responsible for extracting features from the input data and performing complex calculations.

With the development of deep learning, it has been found that increasing the number of model parameters and layers can improve the model's expression ability and prediction performance, leading to the concept of large models. Large models usually consist of billions or tens of billions of parameters, which allows them to have higher accuracy and generalization ability, leading to significant breakthroughs in many complex tasks.

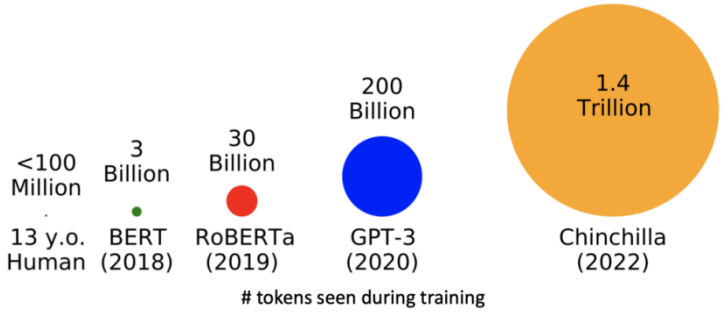

Fig. 1 Number of tokens in large language model(LLM)

2.2 Achievements of Large Models

Large models have achieved many important achievements in fields such as natural language processing, image recognition, and speech recognition. Here are several representative achievements:

1.Natural language processing: BERT is a large pre-trained language model based on the Transformer architecture. It can achieve excellent performance in various natural language processing tasks, such as question answering, text classification, and named entity recognition, through pre-training and fine-tuning. The success of BERT indicates the important application prospects of large models in the field of natural language processing.

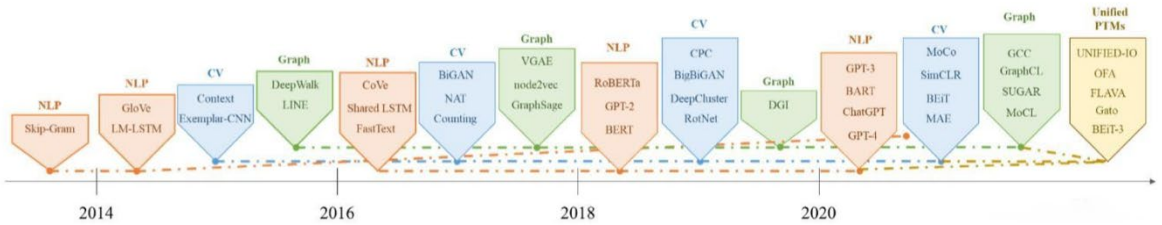

Fig. 2 The evolutionary path of natural language models

2.Image recognition:ResNet is a deep convolutional neural network based on residual connections, which can solve the problems of gradient disappearance and overfitting caused by increasing model depth. ResNet has achieved results surpassing human level on the ImageNet dataset, and has become one of the classic models in the field of image recognition.

3. Speech recognition: DeepSpeech is a deep learning-based speech recognition model that can directly extract text information from speech waveforms. DeepSpeech has achieved excellent results in English speech recognition tasks and can be easily extended to other languages and dialects.

As an important development direction of deep learning technology, large models have become an important driving force for the development of artificial intelligence. The development of large

models has gone through multiple stages of evolution, involving breakthroughs in many key technologies and algorithms, and has also benefited from the development of computer hardware. Large models have achieved many important achievements in fields such as natural language processing, image recognition, and speech recognition, and there will be more applications and breakthroughs in the future. Therefore, the research and application of large models are of great significance and can promote the development and application of artificial intelligence technology.

2.3 Large-Scale Model Capabilities

The development of large-scale models in AI is driven by the desire to achieve better performance and more advanced capabilities. With the introduction of the Transformer architecture, the field of natural language processing (NLP) was revolutionized, and large-scale models based on this architecture quickly came to dominate the field. These models are trained on massive amounts of unlabelled text data, and the amount of data used for pre-training has been increasing rapidly, leading to exponential growth in the number of parameters in these models. More recently, the Transformer architecture has also been applied to computer vision tasks, with Google's ViT model being the first example of this. This has led to the development of large-scale models that are not limited to language understanding but can also be used for vision and other modalities. Multi-modal models, which combine vision, sound, and other modalities, are expected to push AI to new heights, and the combination of generative AI techniques with multi-modal capabilities has the potential to unleash even more creativity in AI-generated content.

Large models have strong generality and can be fine-tuned for different applications. Microsoft 365 Copilot is a good example, where Microsoft has integrated GPT-4 and Microsoft Graph into Office software such as Word, PowerPoint, Excel, Outlook, and Teams. This integration has significantly improved the level of office intelligence, allowing for automatic document generation, presentation beautification, and data summarization. According to gpt3 demo, as of March 2023, there are 725 and 36 derivative apps based on GPT-3 and GPT-4, respectively, which are widely distributed in various content generation fields such as office, notes, conversation, search, advertising, drawing, and music composition.

3. The Value of GPT

The vitality and creativity exhibited by ChatGPT have caused the world to take notice. ChatGPT is an AI chatbot developed by OpenAI. Since its launch in November 2022, it has gained rapid global attention, with 1 million users reached in just 5 days and over 100 million monthly active users in 2 months. We believe that ChatGPT has received such high attention for two reasons: first, it has a wide reach and the basic version is open to almost everyone for free; second, it has strong technical capabilities, providing a conversational experience close to that of a real human, able to continuously optimize by incorporating context and user feedback, and with empathy reaching the level of a 9-year-old child.

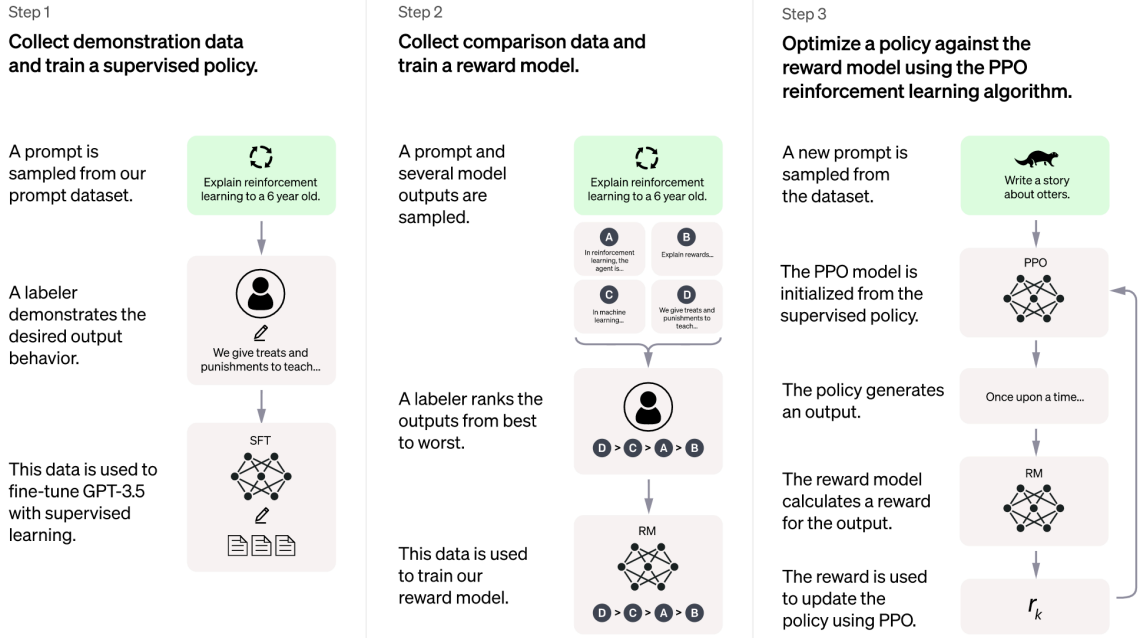

Fig. 3 GPT operating mechanism

The impressive performance of ChatGPT is the result of OpenAI's continuous iteration. In 2017, Google proposed the Transformer neural network architecture, which forms the technical foundation of GPT. Based on the decoder part of Transformer, OpenAI proposed the GPT model (Generative Pre-trained Transformer) and iterated through four major versions from 2018 to 2023, increasing the parameter scale from 117 million for GPT to 175 billion for GPT-3 and continuously improving the model's performance. Based on GPT-3, OpenAI also proposed WebGPT and InstructGPT. The former allows the model to search for information online and cite sources, while the latter introduces the RLHF mechanism (reverse-like feedback mechanism) to output content that meets people's expectations. ChatGPT is the result of the comprehensive application of multiple technological reserves such as GPT-3, WebGPT, and InstructGPT, with a deep technical background.

The emergence of GPT has greatly promoted the development and application of NLP technology. Its performance in natural language generation, machine translation, question-answering systems, text summarization, and dialogue systems has been well validated and applied. In the field of natural language generation, both GPT-2 and GPT-3 have demonstrated powerful language generation capabilities, generating high-quality text, even including poetry and novels. In machine translation, GPT has also performed well, especially in translation tasks between non-English languages. In question-answering systems and text summarization, GPT can also generate high-quality answers and summaries.

When it comes to the impact of GPT models on traditional NLP tasks, we can compare them with traditional rule-based NLP methods. Traditional NLP methods are mainly based on human-made rules and linguistic knowledge, such as syntax and semantic rules, to solve specific NLP tasks. This approach requires a lot of manual work and human design, and often only applies to specific domains and languages, making it difficult to cope with complex and variable natural language.

Compared with traditional rule-based methods, GPT models adopt an end-to-end deep learning approach, training on large amounts of data to automatically learn the rules and patterns of language, thus enabling NLP tasks to be performed in various languages and domains. Taking text classification tasks as an example, compare the performance of GPT models and traditional NLP methods.Compared with traditional rule-based NLP methods, GPT models have better performance in NLP tasks. The reason is that GPT models can automatically learn the rules and patterns of language, eliminating the need for a lot of manual work and human design, thus being better suited to various languages and domains.

Furthermore, we could list out the specific area that GPT model is involved.First, it has greatly improved the performance of language modeling tasks, such as predicting the next word in a sentence or generating coherent text. This is achieved through the use of a large pre-trained language model that has learned to capture the complex and diverse patterns of natural language.Second, GPT has shown promising results in a wide range of downstream NLP tasks, including sentiment analysis, question answering, and language translation. This is due to the transfer learning capabilities of the model, where the pre-trained knowledge can be fine-tuned on specific tasks, resulting in better performance.Lastly, GPT has also tackled the challenge of generating natural and human-like language, which has been a long-standing goal in NLP. The model has demonstrated impressive results in generating high-quality text that is difficult to distinguish from human-generated text.

4. Small-Scale Model

4.1 Theoretical foundation of large models

Large models require more computing resources and longer training times, which may be too expensive or impractical for some downstream businesses. Moreover, small models may already provide sufficient performance for some specific tasks, and the additional advantages of large models may not be worth the investment. Therefore, for businesses with resource constraints or for small-scale tasks, small models may be more suitable. However, large models offer higher accuracy and broader application capabilities. For businesses that require high precision and larger-scale tasks, large models may be more suitable. Over time, the cost of large models is gradually decreasing, making them more attractive in terms of cost-effectiveness. When considering which type of model to use, specific application scenarios, resource constraints, and budgets need to be taken into account, and a balance between cost and performance should be considered.

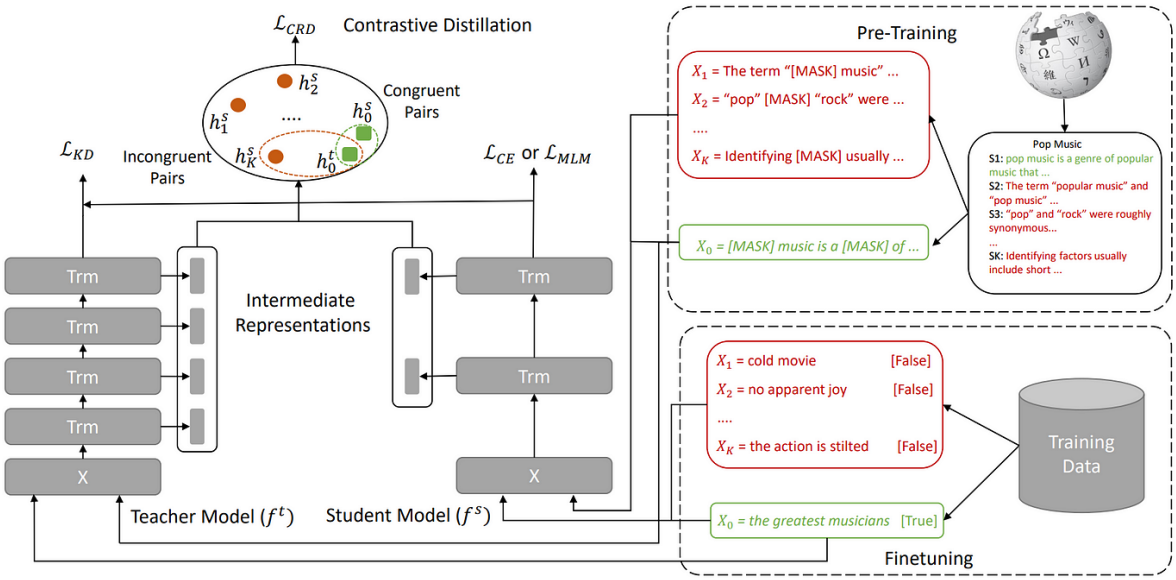

Fig. 4 Principle of Small-Scale Model

Small models are more suitable for meeting the personalized needs of enterprises and allow enterprises to choose suitable models according to their own needs, improving the accuracy and efficiency of the models. The development and application costs of small models are relatively low, making them more easily accepted and applied by small and medium-sized enterprises, promoting the popularization and application of AI technology. Small models do not completely replace big models but serve as a supplement and auxiliary. In some scenarios that require processing large-scale data and complex tasks, big models still have their unique advantages. Small models can achieve lightweight deployment, are tailored to specific use cases, and protect in-house data privacy, making them more suitable for downstream enterprise applications. Industry-specific know-how is a core

competitive advantage that takes a long time to accumulate, and small models are better suited to learning implicit knowledge. Finally, the increasing awareness of data protection in the industry limits the landing of large models, making developing small AI models in-house a preferred option for manufacturers.

4.2 Future Relationship between "Small Models" and "Big Models"

As artificial intelligence technology continues to develop and popularize, the relationship between large models and small models is also receiving increasing attention. The relationship between large and small models in the future will be a dynamic balance process that needs to be analyzed and considered from multiple perspectives.

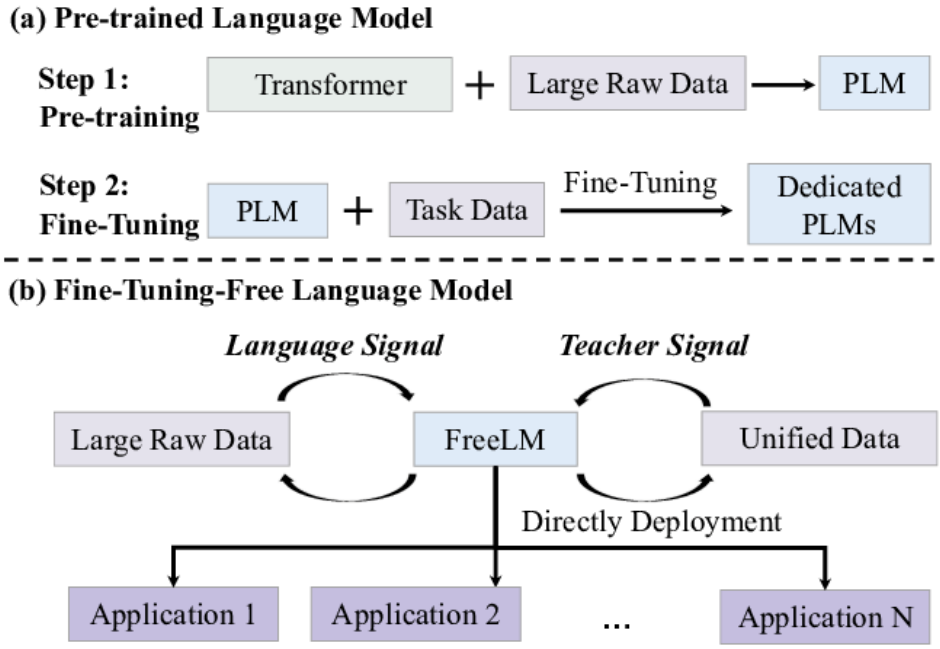

Fig. 5 Brief architecture of the pre-trained model

Firstly, from the perspective of the size and complexity of the model, large models usually have higher accuracy and generalization ability, better adapted to complex tasks and data. However, they also require more computing resources and storage space, and have slower training and inference speed, and higher hardware requirements. In comparison, small models have faster speed, lower latency, and less resource consumption, making them more suitable for use in resource-limited environments. Secondly, from the perspective of task requirements and application scenarios, large and small models also have their own characteristics and advantages. In scenarios where real-time performance and resource consumption are high, such as mobile devices and the Internet of Things, small models are usually more advantageous. While for handling more complex and large-scale tasks, such as natural language processing, computer vision, and speech recognition, large models will be more advantageous. In addition, there are also some contradictions and challenges between large and small models. For example, large models usually require more data and regularization methods for optimization, and are more prone to overfitting and other issues; while the accuracy and performance of small models may be limited and require more efficient algorithms and model designs for optimization.

5. Model to Product

The application of large and small artificial intelligence models in various industries has become one of the hottest topics today. These models not only provide more accurate predictions and decisions but can also significantly reduce enterprise costs and improve efficiency. In this paper, we

will explore the applications of large and small artificial intelligence models in various industries and analyze how they meet user needs.

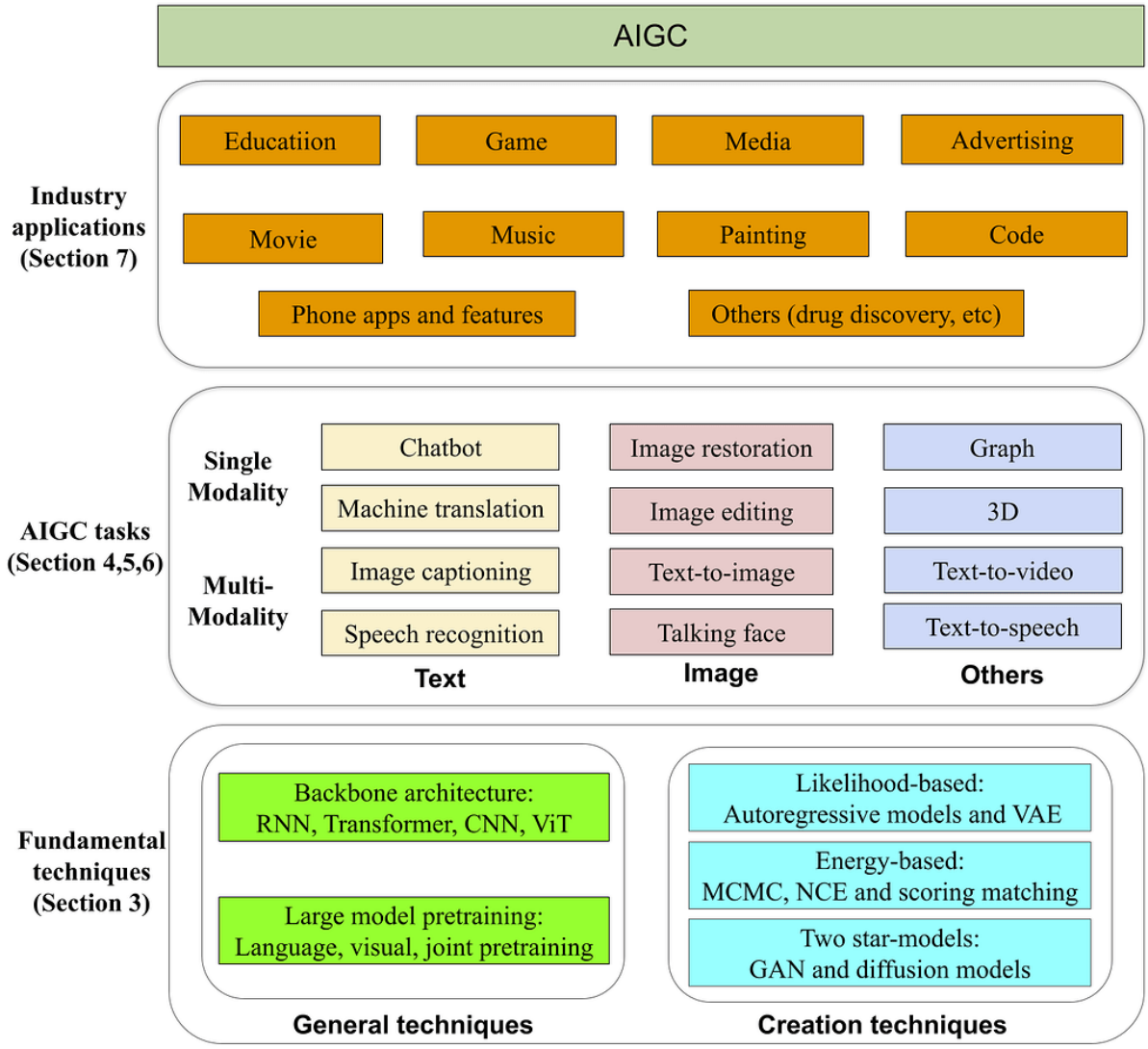

Fig. 6 AIGC categories

When it comes to the applications of large artificial intelligence models in various industries. Large artificial intelligence models typically have millions to billions of parameters and require a lot of computing resources for training and inference. This enables them to perform well on various tasks such as natural language processing, image recognition, and speech recognition. In the banking industry, large artificial intelligence models can be used for automated risk assessment and fraud detection. By analyzing customer transaction history and behavior patterns, these models can generate accurate credit scores and help banks identify potentially fraudulent transactions. In the healthcare industry, large artificial intelligence models can be used for disease diagnosis and drug development. These models can analyze large amounts of medical records to identify patients' symptoms and the best treatment methods. Additionally, these models can help pharmaceutical companies predict the effects and side effects of new drugs. In the retail industry, large artificial intelligence models can be used for recommendation systems and personalized marketing. By analyzing customer purchase history and preferences, these models can generate accurate recommendation lists and help retailers increase sales. Because these models require a lot of computing resources and data, their deployment and use are costly, and they need to handle a large amount of data, which raises privacy and security concerns.

When it comes to the applications of small artificial intelligence models in various industries. Small artificial intelligence models typically have thousands to millions of parameters and can run on resource-constrained devices such as smartphones and IoT devices. This enables them to perform

well on various edge devices such as speech recognition, image classification, and recommendation systems. In the smart home industry, small artificial intelligence models can be used for voice control and smart appliances. These models can run on smart speakers and smartphones to recognize users' voice commands and control home appliances. In the logistics and supply chain industry, small artificial intelligence models can be used for logistics monitoring and smart scheduling. These models can run on IoT devices to monitor the location and status of goods in real-time and intelligently allocate transportation resources. In the human-computer interaction field, small artificial intelligence models can be used for emotion recognition and gesture recognition. These models can run on smartphones and smartwatches to recognize users' emotions and gestures, providing a more intelligent user experience. Because these models have fewer parameters and computing power, their performance may be limited, and a trade-off between model size and performance may be nccecessary.

6. Commercial AI and AI Product

6.1 Commercial Prerequisite

The prerequisites for commercializing AI products are multifaceted. First and foremost, there must be sufficient data support, as AI learning and training are based on a large amount of data. Secondly, appropriate algorithms and models are needed to solve specific problems, which require constant research and development. At the same time, product design needs to meet user needs and be able to satisfy their actual requirements in order to enhance market competitiveness. In addition, the business and profit models also need to be fully considered to ensure that the product can generate profit and sustain development. Moreover, compliance with legal regulations and other related issues also require attention to ensure the product is legal, safe, and reliable. Lastly, team building and talent reserves are also crucial, and a professional, efficient, and innovative team is needed to provide full support for product research and promotion. In summary, data, algorithms, user needs, business models, legal regulations, team building, and other factors are all essential prerequisites for commercializing AI products.

6.2 AI product

Throughout human history, every technological breakthrough has corresponded to structural changes. Today, we have developed PC internet and mobile internet, which have shifted the center of human industrial structure from the primary and secondary industries to the tertiary industry, with countless marginal costs transformed into fixed costs behind it. The efficacy of artificial intelligence models will also correspond to the fixation of vague marginal costs such as information retrieval and decision-making.

From the perspective of user needs, the development of artificial intelligence products can be considered from these two perspectives: whether users' real needs can be well identified, and whether the way products meet these needs is good enough. For the first perspective, product developers need to fully understand users' needs through market research, user feedback, and other means, rather than simply imposing their own ideas on users. At the same time, it is also necessary to realize that users' needs are diverse, and different users have different needs, so it is necessary to develop different products or provide different services based on different users' needs.For the second perspective, product developers need to consider how to efficiently meet users' needs through artificial intelligence technology. Artificial intelligence technology can help products better identify users' needs, improve user experience and product efficiency. At the same time, product developers also need to pay attention to fixing users' marginal costs, such as reducing user operation steps and providing more intuitive operation interfaces.

To maximize the effectiveness and return on investment of AI products, AI product developers must take a user-centric approach, conduct user research and continuous feedback collection, optimize product quality and performance, control product functionality and performance based on market demand, and conduct market and data analysis. Furthermore, AI product developers should

continuously understand user needs and expectations and formulate target thresholds based on the product's actual situation. By following these strategies, AI product developers can ensure their products meet user and market demand and maximize their ROI.

6.3 The Opportunities of AI

There are three major online consumer demands. Firstly, the demand for online shopping platforms is growing rapidly. With the rapid development of the internet, more and more people are using electronic devices to purchase goods. Online shopping is convenient, fast, and diverse, saving users time and energy. Online shopping platforms should provide clear and understandable product information and user evaluations, so that users can make informed purchase decisions. In addition, they should provide secure, convenient payment methods and after-sales service to enhance users' shopping experience and trust.Secondly, the demand for online entertainment is also huge. In the world of the internet, people can easily enjoy various entertainment activities such as movies, TV shows, music, and games at home. Online entertainment platforms should provide diversified content choices and high-quality services to attract and retain users. Moreover, online entertainment platforms also need to provide convenient user experiences and personalized recommendation functions to enhance users' participation and stickiness.Finally, the demand for online learning and education is also increasing. Online learning platforms should provide high-quality course content and a good learning experience to meet users' needs and expectations. In addition, online learning platforms also need to provide personalized learning plans and educational resources to meet the needs and levels of different users.

7. Conclusion

In conclusion, the rapid development of large AI models presents exciting opportunities for the field of AI. As these models become more powerful, they offer the potential for more sophisticated and effective AI products. However, it is important to remember that AI products must ultimately serve the needs of users. This means taking a user-centric approach to the development and commercialization of AI products, and ensuring that these products are designed to meet the needs of specific user groups. By doing so, AI developers and businesses can create AI products that are both successful and impactful.

References

[1] Brown, T. B., et al. Language models are few-shot learners. Advances in neural information processing systems. Vol. 33 (2020).

[2] Radford, A., et al. Language models are unsupervised multitask learners. OpenAI blog. Vol. 1 (2019) No. 8, p. 9.

[3] D. Silver, J. Schrittwieser, K. Simonyan, et al. Mastering the game of Go without human knowledge. Nature. Vol. 550 (2017) No. 7676, p. 354-359.

[4] Zhang, X., et al. A survey of recent advances in unsupervised representation learning for natural language processing. Journal of Big Data. Vol. 9 (2022) No. 1, p. 1-31.

[5] Amodei, D., et al. AI and compute. arXiv preprint arXiv:1804.08328.

[6] Radford, A., et al. Improving language understanding by generative pre-training. .

[7] Wu, J., et al. Beyond accuracy: Behavioral testing of NLP models with CheckList. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Date and location not specified, p. 4908-4922.

[8] Radford, A., et al. Language models are unsupervised multitask learners. OpenAI blog. Vol. 1 (2019) No. 8, p. 9.

[9] Li, Y., et al. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. arXiv preprint arXiv:1908.05969.

[10] Liu, Y., et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Date and location not specified, p. 6657-6666.

[11] Zhang, Z., et al. ChineseNER: An integrated and practical named entity recognition system for Chinese social media text. Information Processing & Management. Vol. 57 (2020) No. 3, p. 102098.

[12] Zhang, Y., et al. The emergence of small AI models. Nature Machine Intelligence. Vol. 3 (2021) No. 5, p. 331-335..

[13] Sze, V., et al. Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE. Vol. 105 (2017) No. 12, p. 2295-2329

[14] J. Howard, S. Ruder. Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, 2018, p. 328-339.

[15] D. Silver, A. Huang, C. J. Maddison, et al. Mastering the game of Go with deep neural networks and tree search. Nature. Vol. 529 (2016) No. 7587, p. 484-489.

[16] Deng, L., et al. Deep learning: Methods and applications. Foundations and Trends in Signal Processing. Vol. 7 (2014) No. 3-4, p. 197-387.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. Nature. Vol. 518 (2015) No. 7540, p. 529-533.

[18] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. Advances in neural information processing systems. Long Beach, 2017, p. 5998-6008.

[19] R. Collobert, J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning. Helsinki, 2008, p. 160-167.

[20] Lewis, M., et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Date and location not specified, p. 7871-7880.

[21] Radford, A., et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

[22] Zhang, Y., et al. Scaling up natural language processing through data-efficient learning. arXiv preprint arXiv:2101.00027.