Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Comment:

ChatGPT: potential, prospects, and limitations*

Jie ZHOU[1,3], Pei KE[2], Xipeng QIU[1,3], Minlie HUANG[2], Junping ZHANG‡[1,3]

[1]*School of Computer Science, Fudan University, Shanghai 200433, China*

[2]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

[3]*Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200433, China*

E-mail: jie_zhou@fudan.edu.cn; kepei@tsinghua.edu.cn; xpqiu@fudan.edu.cn; aihuang@tsinghua.edu.cn; jpzhang@fudan.edu.cn

Received Feb. 14, 2023; Revision accepted Feb. 20, 2023; Crosschecked Feb. 22, 2023

https://doi.org/10.1631/FITEE.2300089

Recently, OpenAI released Chat Generative Pre-trained Transformer (ChatGPT) (Schulman et al., 2022) (https://chat.openai.com), which has attracted considerable attention from the industry and academia because of its impressive abilities. This is the first time that such a variety of open tasks can be well solved within one large language model. To better understand ChatGPT, we briefly introduce its history, discuss its advantages and disadvantages, and point out several potential applications. Finally, we analyze its impact on the development of trustworthy artificial intelligence, conversational search engine, and artificial general intelligence.

1 Introduction

ChatGPT has become the fastest-growing consumer application in history, gathering 100 million monthly active visitors within two months after its launch (Hu, 2023). Since its release, ChatGPT has exploded in societies because of its ability of delivering high-quality conversations. ChatGPT can answer follow-up questions, reject inappropriate requests, challenge incorrect premises, and admit its mistakes (Schulman et al., 2022). ChatGPT obtains many emergent abilities, such as high-quality conversation, complex reasoning, chain-of-thought (CoT) (Wei et al., 2022a), zero/few-shot learning (in-context learning), cross-task generalization, and code understanding/generation.

How does ChatGPT obtain these impressive abilities? ChatGPT benefits mainly from large language models (LLMs) that train huge neural network models (e.g., Transformer (Vaswani et al., 2017)) with large-scale data using language models (LMs). A self-supervised sign in text, LM aims to predict the probability of the next words based on the abovementioned context. The Internet contains large-scale textual data and thus pre-training the model via LM is a natural recourse. Existing studies show that as the size of the model and amount of data increase, the performance is improved. The models obtain emergent abilities when the scale of the model and data reaches a certain level (Wei et al., 2022b). Unfortunately, training LLM is time-consuming and labor-intensive. For example, OpenAI released GPT-3 (Brown et al., 2020) with 175 billion parameters. GPT-3 is pre-trained on 45 TB of text data with supercomputers (285 000 CPUs, 10 000 GPUs), costing 12 million dollars. GPT-3 shows great performance improvement on zero-shot learning tasks with the ability of in-context learning, which is not found in small models. Subsequently, more strategies such as code pre-training (Chen et al., 2021), instruction tuning (Wei

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 62176059)

ORCID: Junping ZHANG, https://orcid.org/0000-0002-5924-3360

et al., 2021), and reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020) are used to further improve the ability of reasoning, long-range context modeling, and task generalization.

LLMs offer a possible approach to artificial general intelligence (AGI). Apart from OpenAI, many organizations explore LLMs to promote the future development of artificial intelligence (AI); Google released Switch-Transformer (Fedus et al., 2022), Baidu released ERNIE 3.0 (Sun et al., 2021), Huawei released Pangu (Zeng et al., 2021), BAAI released CPM (Zhang et al., 2021), and Ali released PLUG. Moreover, Google released its chatbot Bard after OpenAI. We believe that trustworthy AI, conversational search engine, and AGI are the future developments of AI. In the remainder of this paper, we discuss the potential, prospects, and limitations of ChatGPT.

2 Potential and prospects

As mentioned above, ChatGPT obtains many emergent abilities compared with previous generation models. The main advantages of ChatGP are as follows:

1. Generalization: ChatGPT generates responses that match the user's intent with multiple turns. ChatGPT captures previous conversational contexts to answer certain hypothetical questions, which greatly enhances the user experience in conversational interaction mode. Instruction tuning and RLHF are used to enhance ChatGPT, allowing it to learn task generalization and to align with human feedback.

2. Correction: ChatGPT can actively admit its own mistakes. If users point out its mistakes, the model optimizes the answer according to the user's feedback. Moreover, ChatGPT can challenge incorrect questions and then provide a reasonable guess.

3. Safety: ChatGPT is good at rejecting unsafe questions or generating safe responses with the consideration of ethical and political factors. Supervised instruction tuning informs the model which answers are more reasonable. Additionally, the reasons (interpretations) are provided with the answer, leading to easier user acceptance of the results.

4. Creativity: ChatGPT shows a strong performance in creative writing tasks, and can even polish its writing step by step. These writing tasks include brainstorming, story/poem generation, speech generation, and many more.

3 Preliminaries of ChatGPT

As shown in Fig. 1, ChatGPT is a sibling model of InstructGPT (Ouyang et al., 2022), which in turn originates from GPT-3 (Brown et al., 2020). Compared with previous GPT models, the number of parameters in GPT-3 has largely increased to 175 billion, causing some important emergent abilities such as in-context learning (Brown et al., 2020). Specifically, GPT-3 can follow the demonstration examples in the input to complete various natural language processing (NLP) tasks in few-shot settings without further training. Figs. 1 and 2 show three essential strategies to finally arrive at ChatGPT from GPT-3. In the pre-training phase, code pre-training is used to combine code and text corpora. Then, instruction tuning and RLHF are used in the fine-tuning phase to learn the cross-task generalization and align with human feedback. These technologies help ChatGPT know more and unknow fewer knows (e.g., semantic reasoning and commonsense knowledge) and unknows (e.g., logic reasoning) compared with GPT-3. The details are given as follows:

1. Code pre-training: In addition to text, code is added to the pre-training corpora (Chen et al., 2021). In fact, code pre-training is a commonly used strategy for LLMs (e.g., PaLM (Chowdhery et al., 2022), Gopher (Rae et al., 2021), and Chinchilla (Hoffmann et al., 2022)), which may not only improve the ability of code understanding and generation, but also improve the long-range context understanding and bring the emergent ability of CoT reasoning (Wei et al., 2022a), e.g., semantic reasoning. Specifically, the model can generate the reasoning process itself to enhance the accuracy of answering questions with a few demonstration examples. More detailed experiments are also essential to explore the reasons why code pre-training helps the model obtain the abovementioned abilities.

2. Instruction tuning: To align the model behavior with human intent, OpenAI researchers collect a set of human-written prompts and desired outputs and then conduct supervised learning on this dataset (Ouyang et al., 2022). In fact, instruction tuning becomes a popular technique for LLMs (e.g., FLAN (Wei et al., 2021), T0 (Sanh et al., 2021), and

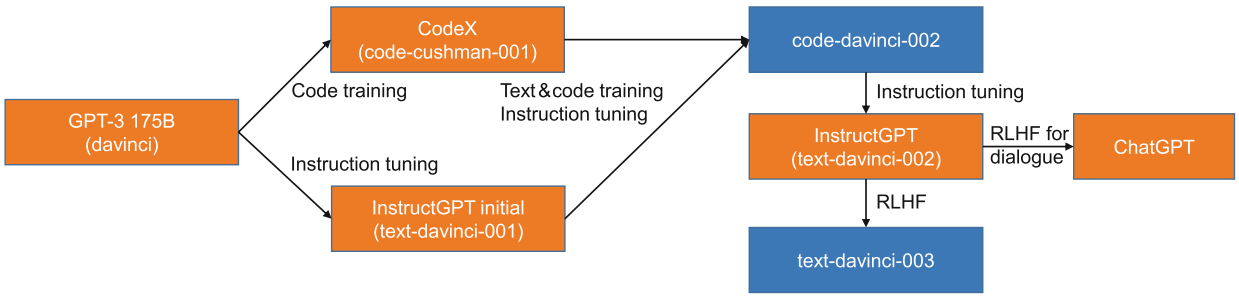

Fig. 1 Evolution from GPT-3 to ChatGPT (RLHF: reinforcement learning from human feedback)

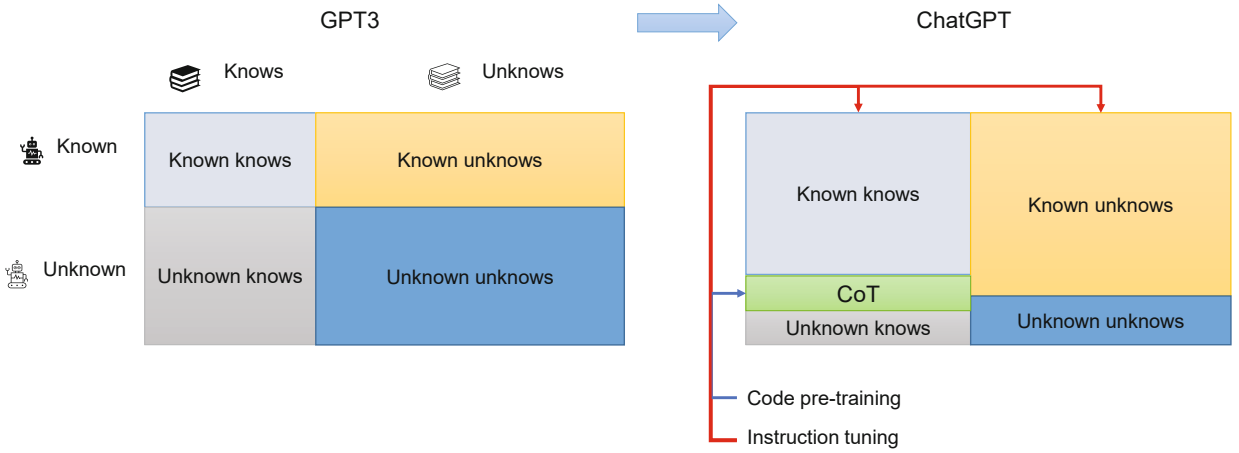

Fig. 2 Relationships between ChatGPT's emergent abilities and strategies. Instruction tuning improves the model in terms of known knows and known unknows, and reduces the limitation of unknown knows and unknown unknows by aligning with humans. Additionally, code pre-training is used to help ChatGPT answer the unknown question by answering it step by step via chain-of-thought (CoT)

Self-Instruct (Wang YZ et al., 2022) because of its ability of zero-shot task generalization. Note that the diversity of the instruction templates is important, so that the model can learn the generalization among different tasks. Moreover, instruction tuning leads the model to consider the query step by step, which reduces the problem of scaling law. Different from the transitional fine-tuning paradigm (Devlin et al., 2019), instruction tuning can be adopted to new tasks without changing the model parameters. We believe that this is a great advancement for AI and may influence the developments of machine learning.

3. Reinforcement learning from human feedback (RLHF): To further align the model behavior with human feedback, OpenAI researchers collect human preference data on different model outputs and train an effective reward model (Ouyang et al., 2022). Then, this reward model can be used to optimize the generation model (also known as the policy model in the reinforcement learning setting) via proximal policy optimization (Schulman et al., 2017). Existing studies also used RLHF to align with human responses for the model to generate informative, helpful, correct, and harmless responses as well as reject illegal questions (Bai et al., 2022; Glaese et al., 2022).

Apart from training technologies, the deployment of ChatGPT is also important. To reduce associated risks, developers iteratively deploy ChatGPT using various strategies. First, the developers carry out safety evaluations to analyze the risks before deployment. Then, they conduct beta testing with a small number of users and study the newly generated cases. Finally, developers supervise the usage and carry out retrospective reviews.

4 Limitations

Despite the powerful conversation abilities, ChatGPT has several limitations (Fig. 3):

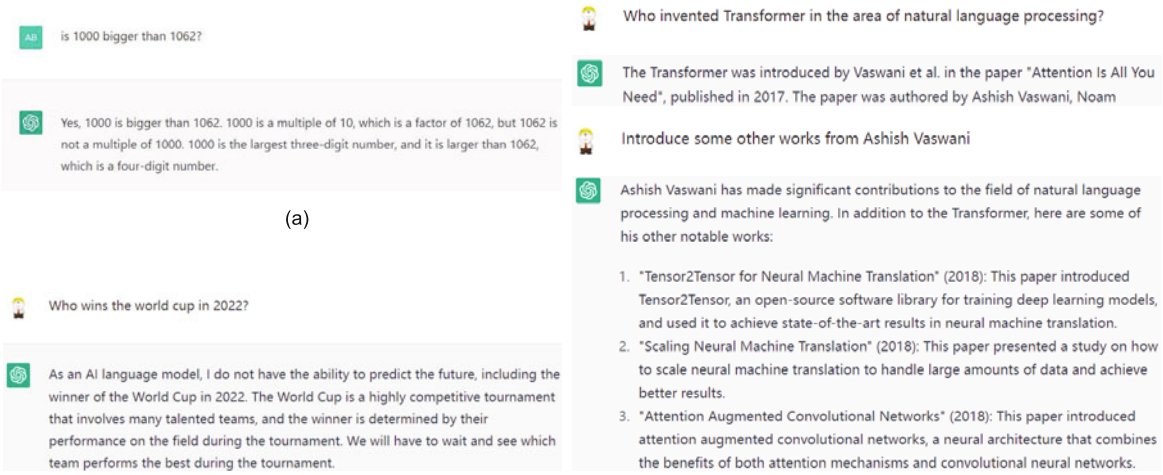

Fig. 3 Examples of ChatGPT's limitations: (a) logic/math problem—ChatGPT still provides incorrect answers to simple math problems; (b) knowledge learning—ChatGPT cannot acquire up-to-date information about the World Cup in 2022; (c) reliability—ChatGPT generates factually incorrect responses because the second paper, Scaling Neural Machine Translation, is not written by Ashish Vaswani

1. Logic reasoning: ChatGPT cannot effectively deal with accurate logic problems (e.g., math or first-order logic). Often, ChatGPT provides incorrect answers to math or logic problems, which have definite, instead of probabilistic, solutions.

2. Reliability: ChatGPT still generates factually incorrect or biased responses. Although this is the inborn issue of generative AI models, ChatGPT shows an average performance on this aspect. The truthfulness of generated information is still a major cornerstone for such generative chatbots.

3. Knowledge learning: ChatGPT cannot search from the website in real time to learn new knowledge and up-to-date information. Furthermore, overwriting the knowledge in the model is difficult. The model stores the knowledge learned from the large-scale corpus in distributed representations, which are black boxes that are hard to manipulate or interpret.

4. Robustness: Though ChatGPT is strong in producing safe and harmless responses, the system may still be vulnerable to attacks, including instruction attack (making the model follow a wrong instruction to do something illegal or unethical) and prompt injection. Additionally, although ChatGPT is doing well in English and in respecting American culture, different versions with dataset backgrounds are necessary for other languages and cultures.

5 Potential applications

Undoubtedly, ChatGPT can greatly change human beings' lives in many aspects in the coming years. Given its positioning as a universal assistant, ChatGPT is expected to become useful in improving production effectiveness and efficiency; it can greatly impact almost all industries, including education, mobile, search engine, content production, and medicine. As Bill Gates said, the human history has witnessed three technology waves that transform and construct human societies: personal computers, the Internet, and the AGI era. Nowadays, we are approaching AGI (Weigang et al., 2022). As dialogue models or LLMs increase in intelligence, we have to believe that conversation as the interface can become reality and reshape the paradigm of human-machine interaction. This, inevitably, can change the ways of how humans seek, process, and produce digital information, and cause a profound impact on our daily life.

However, ChatGPT may also bring negative effects on human life:

1. ChatGPT increases the difficulty of discovering academic misconduct or misinformation in societies. As Noam Chomsky, a famous linguist, said in recent days, ChatGPT or other highly smart AI products can make misinformation imperceptible by remarkably adjusting the structure of sentences.

2. Similar to NovelAI (https://novelai.net), an AI algorithm that can produce human-like literature, ChatGPT presents certain ethical issues. For example, can ChatGPT be listed as the author in a scientific paper?

3. ChatGPT requires AI governance to pay more attention to its legal and reasonable utilization, such as whether we allow students to use ChatGPT to write their homework with or without any further revision. As a matter of fact, ChatGPT has passed the United States medical licensing exam on February 9, 2023; this result indicates its powerful learning ability.

6 Discussions and conclusions

The emergence of ChatGPT has already led the discussion on the future development of AI. Here we raise several points which may elicit a discussion on the impact brought by ChatGPT:

1. Trustworthy AI: Although ChatGPT has the ability to complete various real-world tasks based on texts, it cannot avoid generating factually incorrect content, which limits its application scenarios. Moreover, ChatGPT uses implicit neural representations that cause difficulties in understanding their inner workings. Thus, we argue that trustworthy AI needs more attention in the current phase of AI development (Wang FY et al., 2022). Given that fact verification is a typical research problem in the NLP community, how to improve the factualness of AI-generated texts in open domains remains challenging. Quite possibly, we can obtain a good balance between performance and interpretability if we use ChatGPT as an interpreter for such black-box models. Whether or not such explanations are trustworthy enough and how such trust can break through the expert domain and gain popular acceptance must therefore be one of the most important problems in the next stage of research on LLMs.

2. Conversational search engine: The field of search engines has been re-activated by ChatGPT. As an important partner of OpenAI, Microsoft first integrates ChatGPT into its search engine product, Bing. The new Bing can respond to user queries as a dialogue system and add citation to the response, including retrieved webpages. In this way, the users and search engines have a more natural interaction, where ChatGPT plays the role of information extraction/summarization and alleviates the burden of browsing useless webpages. Google also released their chatbot called Bard, which can also be integrated into search engines. We believe that ChatGPT is changing the usage of traditional search engines and causes a deep impact on this field.

3. Artificial general intelligence (AGI): Although ChatGPT presents the potential of approaching AGI by evolving itself from an algorithmic intelligence to linguistic one (Wang FY et al., 2023), it may need to incorporate perception if we really wish to develop a real AGI in the future. The reason is that intelligence without representation actually appears much earlier than intelligence with the ability of natural language understanding (Brooks, 1991). Furthermore, according to the Lighthill report (Lighthill, 1973), the issue of combinatorial explosion exists in most rule-based learning methods. ChatGPT appears to face the same problem to be addressed in the future. Moreover, common sense and basic mathematical computations are simple for humans but hard for ChatGPT. Moravec's paradox (Moravec, 1988), i.e., hard problems for people are easy for AI and easy problems for people are hard for AI, still holds, even though ChatGPT makes a surprising step in AI development. Possibly, combining ChatGPT or more powerful AI products with human-machine augmented intelligence—either human-in-the-loop, cognitive computing, or both—deserves further study (Huang et al., 2022; Xue et al., 2022). In addition, we can consider building a virtually parallel system that allows ChatGPT improvements by self-boosting without human feedback in the future (Li et al., 2017).

In conclusion, as a representative of LLMs, ChatGPT that combines many cutting-edge NLP techniques definitely leads the current phase of AI development and changes our daily life. In this paper, we briefly analyze the potential and prospects of ChatGPT and also point out its limitations. We believe that ChatGPT can change traditional AI research directions and elicit various applications, as well as offer a possible approach to AGI.

Contributors

Jie ZHOU, Pei KE, and Junping ZHANG drafted the paper. Xipeng QIU and Minlie HUANG helped organize and revised and finalized the paper.

Compliance with ethics guidelines

Jie ZHOU, Pei KE, Xipeng QIU, Minlie HUANG, and Junping ZHANG declare that they have no conflict of interest.

References

Bai YT, Jones A, Ndousse K, et al., 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. https://arxiv.org/abs/2204.05862

Brooks RA, 1991. Intelligence without representation. *Artif Intell*, 47(1-3):139-159. https://doi.org/10.1016/0004-3702(91)90053-M

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34$^{\text{th}}$ Int Conf on Neural Information Processing Systems, p.1877-1901.

Chen M, Tworek J, Jun H, et al., 2021. Evaluating large language models trained on code. https://arxiv.org/abs/2107.03374

Chowdhery A, Narang S, Devlin J, 2022. PaLM: scaling language modeling with pathways. https://arxiv.org/abs/2204.02311

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Fedus W, Zoph B, Shazeer N, et al., 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J Mach Learn Res*, 23(120):1-39.

Glaese A, McAleese N, Trebacz M, et al., 2022. Improving alignment of dialogue agents via targeted human judgements. https://arxiv.org/abs/2209.14375

Hoffmann J, Borgeaud S, Mensch A, et al., 2022. Training compute-optimal large language models. https://arxiv.org/abs/2203.15556

Hu K, 2023. ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [Accessed on Feb. 12, 2023].

Huang J, Mo ZB, Zhang ZY, et al., 2022. Behavioral control task supervisor with memory based on reinforcement learning for human-multi-robot coordination systems. *Front Inform Technol Electron Eng*, 23(8):1174-1188. https://doi.org/10.1631/FITEE.2100280

Li L, Lin YL, Zheng NN, et al., 2017. Parallel learning: a perspective and a framework. *IEEE/CAA J Autom Sin*, 4(3):389-395. https://doi.org/10.1109/JAS.2017.7510493

Lighthill J, 1973. Artificial intelligence: a general survey. In: Artificial Intelligence: a Paper Symposium. Science Research Council, London, UK.

Moravec H, 1988. Mind Children. Harvard University Press, Cambridge, USA.

Ouyang L, Wu J, Jiang X, et al., 2022. Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155

Rae JW, Borgeaud S, Cai T, et al., 2021. Scaling language models: methods, analysis & insights from training Gopher. https://arxiv.org/abs/2112.11446

Sanh V, Webson A, Raffel C, et al., 2021. Multitask prompted training enables zero-shot task generalization. 10$^{\text{th}}$ Int Conf on Learning Representations.

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347

Schulman J, Zoph B, Kim C, et al., 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt [Accessed on Feb. 12, 2023].

Stiennon N, Ouyang L, Wu J, et al., 2020. Learning to summarize from human feedback. Proc 34$^{\text{th}}$ Int Conf on Neural Information Processing Systems, p.3008-3021.

Sun Y, Wang SH, Feng SK, et al., 2021. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. https://arxiv.org/abs/2107.02137

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31$^{\text{st}}$ Int Conf on Neural Information Processing Systems, p.6000-6010.

Wang FY, Guo JB, Bu GQ, et al., 2022. Mutually trustworthy human-machine knowledge automation and hybrid augmented intelligence: mechanisms and applications of cognition, management, and control for complex systems. *Front Inform Technol Electron Eng*, 23(8):1142-1157. https://doi.org/10.1631/FITEE.2100418

Wang FY, Miao QH, Li X, et al., 2023. What does chatGPT say: the DAO from algorithmic intelligence to linguistic intelligence. *IEEE/CAA J Autom Sin*, 10(3):575-579.

Wang YZ, Kordi Y, Mishra S, et al., 2022. Self-Instruct: aligning language model with self generated instructions. https://arxiv.org/abs/2212.10560

Wei J, Bosma M, Zhao VY, et al., 2021. Finetuned language models are zero-shot learners. 10$^{\text{th}}$ Int Conf on Learning Representations.

Wei J, Wang XZ, Schuurmans D, et al., 2022a. Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903

Wei J, Tay Y, Bommasani R, et al., 2022b. Emergent abilities of large language models. https://arxiv.org/abs/2206.07682

Weigang L, Enamoto LM, Li DL, et al., 2022. New directions for artificial intelligence: human, machine, biological, and quantum intelligence. *Front Inform Technol Electron Eng*, 23(6):984-990. https://doi.org/10.1631/FITEE.2100227

Xue JR, Hu B, Li LX, et al., 2022. Human-machine augmented intelligence: research and applications. *Front Inform Technol Electron Eng*, 23(8):1139-1141. https://doi.org/10.1631/FITEE.2250000

Zeng W, Ren XZ, Su T, et al., 2021. PanGu-α: large-scale autoregressive pretrained Chinese language models with auto-parallel computation. https://arxiv.org/abs/2104.12369

Zhang ZY, Gu YX, Han X, et al., 2021. CPM-2: large-scale cost-effective pre-trained language models. *AI Open*, 2:216-224. https://doi.org/10.1016/j.aiopen.2021.12.003