The significance of GPT4.5 and recursive self-improvement of AI

The release of GPT4.5 has caused people to worry about the bottleneck of pre-training. Many researchers, including Ilya, have predicted that pre-training will encounter bottlenecks when Internet data is gradually exhausted. However, due to the emergence of inference models, such as the earliest OpenAI O1, the later Deepseek R1, and the later Grok3 and Claude 3.7, people have greatly improved the reasoning ability of the model by using post-training techniques such as fine-tuning (SFT) enhanced learning RL.

After the reasoning ability of large models is improved, people and AI can work together, communicate, and research together, just like AlphaFold cooperated with humans to predict protein folding and won the Nobel Prize in Chemistry, which shows that higher-quality new data can be generated. When people feed these high-quality new data to the basic model for pretraining, they can continue to improve the ability of the basic model. Then, the reasoning ability is further improved through post-training. This cycle of interaction continues until the limit of human ability is reached. Then AI can be allowed to communicate with itself to improve, just like alpha_zero. By then, AI can automatically deduce, summarize, associate, and think like humans, even surpassing the human intelligence level, and generate intelligence and data beyond the limits of human capabilities.

This cycle can be called "recursive self-improvement" and "human-machine collaborative innovation". In theory, humans and AI conduct research together and continuously feed new data back to the basic model to improve model performance and reasoning ability.

Therefore, the launch of GPT4.5, although the reasoning ability is not as good as the reasoning model due to the lack of post-inference training, the basic model itself has been improved. If the same post-inference training (fine-tuning) process is carried out, the reasoning ability of the large model will be higher than before.

The reasons include:

1 Better representation ability: After the basic model is improved, its representation and abstraction ability in the data is enhanced, which provides higher quality feature representation for subsequent reasoning training, so that the model can better capture complex patterns and logical relationships.

2 Deeper knowledge integration: The improved basic model can usually integrate more cross-domain and deep knowledge, which helps to integrate and deduce information more accurately and efficiently in the reasoning process.

3 Improvement of training efficiency and generalization ability: The improved model may be easier to converge when facing the same reasoning task, and can be better generalized to new tasks and new scenarios, further improving the overall reasoning ability.

Of course, there are also some challenges in feasibility:

1 Data quality and feedback mechanism: Model improvement depends heavily on the quality of feedback data. If there is bias, noise or erroneous information in the data, it may cause instability or even degradation of model performance.

2 Model autonomy and security: When AI starts autonomous dialogue and self-interpretation, how to ensure that its output is within a controllable range and avoid unforeseen or harmful behavior is an urgent problem to be solved.

3 Technical bottleneck: At present, although AI has made progress in automatic deduction, induction and association, it is still a long way from fully simulating human thinking, and related technologies need to be broken through.

4 Ethics and supervision: If AI continues to evolve itself, how to maintain the ultimate control of humans and ensure that the entire process complies with ethical norms and safety standards is also an important factor that must be considered.

This is a path worth exploring. GPT4.5 is the last generation of pre-training models released publicly. It actually means that the base model will no longer be a product but a basis for post-training. The high price is also to prevent distillation. Altman and Musk and others have asserted that AGI may appear in a few years. There is no wall.