2023 text-generating language model

Generative Pre-trained Transformer 4 (GPT-4) is a multimodal large language model created by OpenAI, and the fourth in its series of GPT foundation models. It was launched on March 14, 2023, and made publicly available via the paid chatbot product ChatGPT Plus, via OpenAI's API, and via the free chatbot Microsoft Copilot. As a transformer-based model, GPT-4 uses a paradigm where pre-training using both public data and "data licensed from third-party providers" is used to predict the next token. After this step, the model was then fine-tuned with reinforcement learning feedback from humans and AI for human alignment and policy compliance.

Observers reported that the iteration of ChatGPT using GPT-4 was an improvement on the previous iteration based on GPT-3.5, with the caveat that GPT-4 retains some of the problems with earlier revisions. GPT-4, equipped with vision capabilities (GPT-4V), is capable of taking images as input on ChatGPT. OpenAI has declined to reveal various technical details and statistics about GPT-4, such as the precise size of the model.

Further information: GPT-3 § Background, and GPT-2 § Background

OpenAI introduced the first GPT model (GPT-1) in 2018, publishing a paper called "Improving Language Understanding by Generative Pre-Training." It was based on the transformer architecture and trained on a large corpus of books. The next year, they introduced GPT-2, a larger model that could generate coherent text. In 2020, they introduced GPT-3, a model with 100 times as many parameters as GPT-2, that could perform various tasks with few examples. GPT-3 was further improved into GPT-3.5, which was used to create the chatbot product ChatGPT.

Rumors claim that GPT-4 has 1.76 trillion parameters, which was first estimated by the speed it was running and by George Hotz.

OpenAI stated that GPT-4 is "more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5." They produced two versions of GPT-4, with context windows of 8,192 and 32,768 tokens, a significant improvement over GPT-3.5 and GPT-3, which were limited to 4,096 and 2,049 tokens respectively. Some of the capabilities of GPT-4 were predicted by OpenAI before training it, although other capabilities remained hard to predict due to breaks in downstream scaling laws. Unlike its predecessors, GPT-4 is a multimodal model: it can take images as well as text as input; this gives it the ability to describe the humor in unusual images, summarize text from screenshots, and answer exam questions that contain diagrams. It can now interact with users through spoken words and respond to images, allowing for more natural conversations and the ability to provide suggestions or answers based on photo uploads.

To gain further control over GPT-4, OpenAI introduced the "system message", a directive in natural language given to GPT-4 in order to specify its tone of voice and task. For example, the system message can instruct the model to "be a Shakespearean pirate", in which case it will respond in rhyming,