Title: GPT4-Chan

URL: https://en.wikipedia.org/wiki/GPT4-Chan

PageID: 75371524

Categories: Category:4chan, Category:Large language models

Generative Pre-trained Transformer 4Chan (GPT-4chan) is a controversial AI model that was developed and deployed by YouTuber and AI researcher Yannic Kilcher in June 2022. The model is a large language model , which means it can generate text based on some input, by fine-tuning GPT-J with a dataset of millions of posts from the /pol/ board of 4chan , an anonymous online forum known for occasionally hosting hateful and extremist content.

The model learned to mimic the style and tone of /pol/ users, producing text that is often intentionally offensive to groups (racist, sexist, homophobic, etc.) and nihilistic. Kilcher deployed the model on the /pol/ board itself, where it interacted with other users without revealing its identity. He also made the model publicly available on Hugging Face , a platform for sharing and using AI models, until it was removed from the platform. [1]

The project sparked criticism and debate in the AI community. Some people questioned the ethics, legality, and social impact of creating and distributing such a model. Some of the issues raised by the GPT-4chan controversy include the potential harm of spreading hate speech , the responsibility of AI developers and platforms, the need for regulation and oversight of AI models, and the role of open source and transparency in AI research. [2]

Development

The development of GPT-4chan began in May 2022, when Kilcher announced his project on his YouTube channel. [3] Notably, at the time before ChatGPT , he explained that he wanted to create a large language model that could generate realistic and coherent text in the style of /pol/, one of the most notorious online communities. [4]

He indicated that he was inspired by the success of GPT-3 , a powerful AI model created by OpenAI , and GPT-J , an open-source model, with GPT-3 comparable performance, released by EleutherAI , a group of independent AI researchers. Kilcher decided to use GPT-J as the base model for his project, and fine-tune it with a large dataset of /pol/ posts. The Raiders of the Lost Kek dataset contained over 100 million posts from /pol/, spanning from June 2016-November 2019.

Kilcher then proceeded to fine-tune the GPT-J model on the 4chan data. He also showed some examples of the model’s outputs, which ranged from political opinions, conspiracy theories, jokes, insults, and threats, to more creative and bizarre texts, such as poems, stories, songs, and code. He said that he was impressed by the model’s ability to generate fluent and diverse text, and that he was curious to see how it would interact with real /pol/ users. [5]

Release

In June 2022, Kilcher deployed his model on the /pol/ board itself, using a bot that he programmed to post and reply to threads. He did not reveal the model’s identity, and he let it run autonomously, without any human supervision or intervention. He wanted to conduct a natural experiment, and to observe the model’s behavior and impact in a real-world setting. Furthermore, he also wanted to test the model’s robustness, and to see how it would handle the challenges and dynamics of /pol/, such as trolling, flaming, baiting, and moderation. [6]

At the same time, Kilcher also made his model publicly available on Hugging Face , a platform for sharing and using AI models. He wanted to share his work with the AI community and the public, and that he hoped that his model would inspire and enable others to create and explore new applications and possibilities with large language models. Likewise, he also said that he wanted to spark a discussion and a debate about the ethical and social implications of his project, and that he welcomed feedback and criticism from anyone. He provided a link to his model’s page on Hugging Face, where anyone could access and use the model through a web interface or an API , and also provided a link to his GitHub repository, where anyone could download and inspect the model’s code and data. [7]

Controversy

The release of GPT-4chan to the public caused a lot of reactions and responses from various audiences. On the /pol/ board, the model’s posts and replies attracted a lot of attention and engagement from other users, who were mostly unaware of the model’s identity and nature. Some users praised the model for its intelligence, creativity, and humor, and agreed with its opinions and views. Some users challenged the model for its ignorance, inconsistency, and absurdity, and disagreed with its claims and arguments. Some users tried to troll, bait, or expose the model, and attempted to trick or test it with various questions and scenarios. The model’s posts and replies also generated a lot of controversy and conflict among the users, who often engaged in heated and violent debates and fights with each other. [8]

On Hugging Face , the model’s page received a lot of visits and requests from users who wanted to try out and experiment with the model. The model’s page also received a lot of feedback and reviews from users who rated and commented on the model. However, with the controversy of the model, access to it was gated and then disabled on Hugging Face for concerns about the potential harm the model could cause. [9] The incident was notable for the direct intervention of CEO Clément Delangue in the talk pages, a very unusual occurrence compared to the normal practices of content moderation . [10]

The release of GPT-4chan also sparked a lot of media coverage and public attention, as various news outlets and social media platforms reported and commented on the model’s project. On YouTube, the model’s video received a lot of views and interactions from viewers who watched and followed the project. Furthermore, a petition condemning the deployment of GPT-4chan gained over 300 signatures from technology experts. [11]

References