OpenAI transcribes millions of hours of YouTube videos to train GPT-4

In an effort to secure high-quality data to train massive artificial intelligence models, major AI companies like OpenAI, Google, and Meta are now stepping up their use of 'shady' data collection tactics. '.

In an effort to secure high-quality data to train massive artificial intelligence models, major AI companies such as OpenAI, Google and Meta are now promoting the use of 'shady' data collection tactics. '. A recent New York Times report said that OpenAI intentionally transcribed more than a million hours of YouTube videos into data to train its most advanced large language model (LLM): GPT-4.

Accordingly, OpenAI has developed the Whisper audio transcription model, helping the company collect data from YouTube videos. The NY Times reported that OpenAI was fully aware this method could come under scrutiny, but continued to do it anyway because they believed the use was completely legal. Interestingly, Google, the company that owns YouTube, is also accused of being involved in similar activity with its AI models, i.e. directly violating the copyright of video creators.

Agreeing with the NY Times, The Information's report emphasized that OpenAI scraped data from YouTube videos and podcasts to train its two AI systems, and hinted that OpenAI president Greg Brockman also knew and agreed with this approach.

In a recent interview with Bloomberg, YouTube CEO Neil Mohan said that the company's policy "doesn't allow downloading of content like transcripts or bits of video, and that's a clear violation our terms of service". However, when asked whether YouTube data was being 'abused' by OpenAI, the CEO only gave a relatively

vague answer: "I have seen reports that YouTube data may have been misused." used or not. I have no information myself."

The NY Times report further claims that some Google employees were aware of OpenAI's YouTube data copying activities, but they were unable to do anything because the Mountain View company itself used Similar method to train your own AI models. However, Google told The NY Times that it only collects video data after the video creator has given consent.

Google even reportedly "adjusted its privacy policy" in June 2023, "to allow data mining of publicly available Google Docs, Google Maps reviews, and many types of documents online to train the company's AI products".