DeepMind develops SAFE, an AI-based app that can fact-check LLMs

March 29 2024, by Bob Yirka

Credit: CC0 Public Domain

A team of artificial intelligence specialists at Google's DeepMind has

developed an AI-based system called SAFE that can be used to fact check the results of LLMs such as ChatGPT. The group has published a paper describing the new AI system and how well it performed on the *arXiv* preprint server.

Large language models such as ChatGPT have been in the news a lot over the past couple of years—they can write papers, give answers to questions and even solve math problems. But they suffer from one major problem: accuracy. Every result obtained by an LLM must be checked manually to ensure that the results are correct, an attribute that greatly reduces their value.

In this new effort, the researchers at DeepMind created an AI application that can check the results of answers given by LLMs and point out inaccuracies automatically.

One of the main ways that human users of LLMs fact-check results is by investigating AI responses using a search engine such as Google to find appropriate sources for verification. The team at DeepMind took the same approach. They created an LLM that breaks down claims or facts in an answer provided by the original LLM and then used Google Search to find sites that could be used for verification and then compared the two answers to determine accuracy. They call their new system Search-Augmented Factuality Evaluator (SAFE).

To test their system, the research team used it to verify approximately 16,000 facts contained in answers given by several LLMs. They compared their results against human (crowdsourced) fact-checkers and found that SAFE matched the findings of the humans 72% of the time. When testing disagreements between SAFE and the human checkers, the researchers found SAFE to be the one that was correct 76% of the time.

The team at DeepMind has made the code for SAFE available for use by anyone who chooses to take advantage of its capabilities by posting in on the open-source site GitHub.

More information: Jerry Wei et al, Long-form factuality in large language models, *arXiv* (2024). DOI: 10.48550/arxiv.2403.18802

Code release: github.com/google-deepmind/long-form-factuality