Scientists find ChatGPT is inaccurate when answering computer programming questions

May 27 2024, by Bob Yirka

A team of computer scientists at Purdue University has found that the popular LLM, ChatGPT, is wildly inaccurate when responding to computer programming questions. In their paper published as part of the Proceedings of the CHI Conference on Human Factors in Computing Systems, the group describes how they pulled questions from the StackOverflow website and posed them to ChatGPT and then measured its degree of accuracy when responding.

The team also presented their findings at the Conference on Human Factors in Computing Systems (CHI 2024) held May 11–16.

ChatGPT and other LLMs have been in the news a lot recently—since such apps have been made available to the general public, they have become very popular. Unfortunately, along with a treasure trove of useful information included in many of the responses given by such apps, there are a host of inaccuracies. Even more unfortunate is that it is not always clear when the apps are giving answers that are wrong.

In this new study, the team at Purdue noted that many programming students have begun using LLMs to not only help write code for programming assignments, but to answer questions related to programming. As an example, a student could ask ChatGPT, what is the difference between a bubble sort and merge sort, or, more popularly, what is recursion?

To find out how accurate LLMs are in answering such questions, the
research team focused their efforts on just one of them—ChatGPT. To find questions to use for testing the app, the researchers used questions freely available on the StackOverflow website—it is a site that has been built to help programmers learn more about programming by working with others in their field of interest. On one part of the site, users can post questions that will be answered by others who know the answers.

The research team used 517 questions found on the site and then measured how often ChatGPT gave the correct answer. The study primarily involved the use of the GPT-3.5 model available in the free version of ChatGPT for the manual responses to 517 questions and utilized the GPT-3.5-turbo API for the larger automated processing of 2,000 additional questions. Data collection took place in March 2023. Sadly, it was correct just 52% of the time. They also found the answers tended to be more verbose than would be the case when a human expert was asked the same question. Researchers also compared the performance with the GPT-4 model. The GPT-4 model performed slightly better than GPT-3.5 by correctly answering 6 out of 21 randomly selected questions that GPT-3.5 had answered incorrectly, but it still generated a majority of incorrect responses (15 out of 21).

Alarmingly, the team found that user study participants preferred the answers given by ChatGPT 35% of the time. The researchers also found that the same users reading the answers given by ChatGPT quite often did not catch the mistakes that were made—they overlooked wrong answers 39% of the time.
