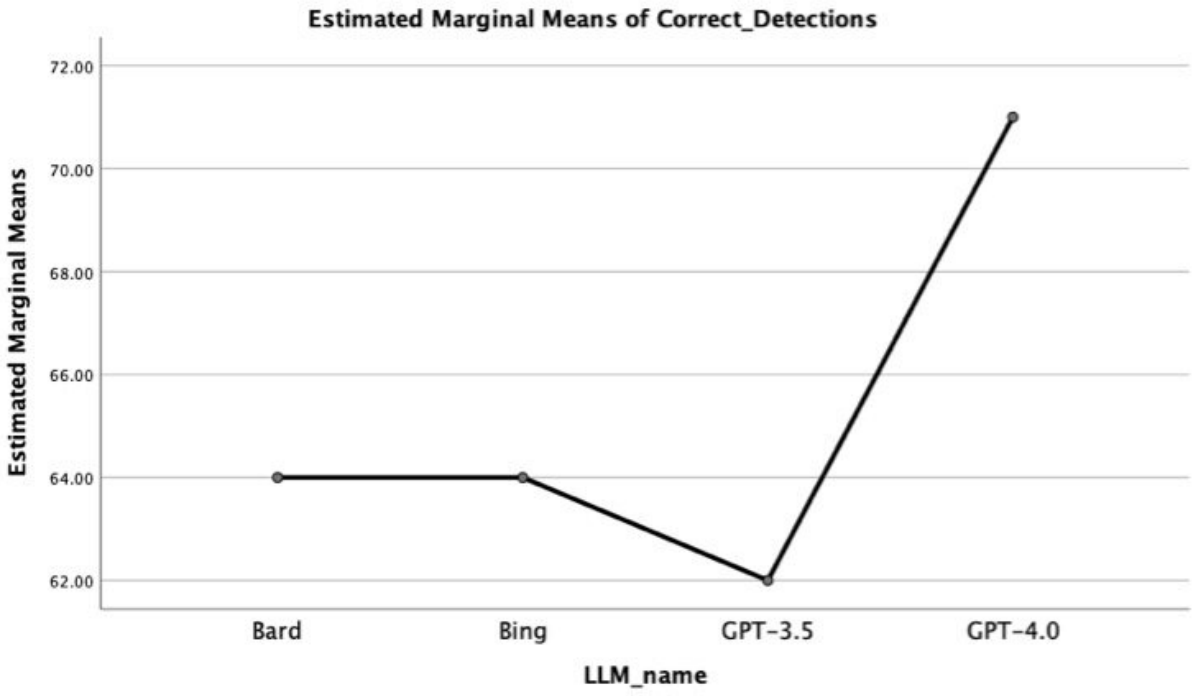
Evaluating the ability of ChatGPT and other large language models to detect fake news

July 17 2023, by Ingrid Fadelli



Graph summarizing the outcome of the study Credit: Kevin Matthe Caramancion

Large language models (LLMs) are an evolution of natural language processing (NLP) techniques that can rapidly generate texts closely resembling those written by humans and complete other simple language-related tasks. These models have become increasingly popular after the public release of Chat GPT, a highly performing LLM developed by

OpenAI.

Recent studies evaluating LLMs have so far primarily tested their ability to create well-written texts, define specific terms, write essays or other documents, and produce effective computer code. Nonetheless, these models could potentially help humans to tackle various other real-world problems, including fake news and misinformation.

Kevin Matthe Caramancion, a researcher at University of Wisconsin-Stout, recently carried out a study evaluating the ability of the most well-known LLMs released to date to detect if a news story is true or fake. His findings, in a paper on the preprint server *arXiv*, offers valuable insight that could contribute to the future use of these sophisticated models to counteract online misinformation.

"The inspiration for my recent paper came from the need to understand the capabilities and limitations of various LLMs in the fight against misinformation," Caramancion told Tech Xplore. "My objective was to rigorously test the proficiency of these models in discerning fact from fabrication, using a controlled simulation and established fact-checking agencies as a benchmark."

"We evaluated the performance of these large language models using a test suite of 100 fact-checked news items from independent fact-checking agencies," Caramancion said. "We presented each of these news items to the models under controlled conditions and then classified their responses into one of three categories: True, False, and Partially True/False. The effectiveness of the models was measured based on how accurately they classified these items compared to the verified facts provided by the independent agencies."

Misinformation has become a crucial challenge in recent decades, as the internet and social media have enabled the increasingly rapid

dissemination of information, irrespective of whether it is true or false. Many computer scientists have thus been trying to devise better fact-checking tools and platforms that allow users to verify news that they read online.

Despite the many fact-checking tools created and tested to date, a widely adopted and reliable model to combat misinformation is still lacking. As part of his study, Caramancion set out to determine whether existing LLMs could effectively tackle this worldwide issue.

He specifically assessed the performance of four LLMs, namely Open AI's Chat GPT-3.0 and Chat GPT-4.0, Google's Bard/LaMDA, and Microsoft's Bing AI. Caramancion fed these models the same news stories, which were previously fact-checked, and then compared their ability to determine if they were true, false, or partly true/false.

"We performed a comparative evaluation of major LLMs in their capacity to differentiate fact from deception," Caramancion said. "We found that OpenAI's GPT-4.0 outperformed the others, hinting at the advancements in newer LLMs. However, all models lagged behind human fact-checkers, emphasizing the irreplaceable value of human cognition. These findings could lead to an increased focus on the development of AI capabilities in the field of fact-checking while ensuring a balanced, symbiotic integration with human skills."

The evaluation carried out by Caramancion showed that ChatGPT 4.0 significantly outperforms other prominent LLMs on fact-checking tasks. Further studies testing LLMs on a wider pool of fake news could help to verify this finding.

The researcher also found that human fact-checkers still outperform all the primary LLMs he assessed. His work highlights the need to improve these models further or combine them with the work of human agents if

they are to be applied to fact-checking.

"My future research plans revolve around studying the progression of AI capabilities, focusing on how we can leverage these advancements while not overlooking the unique cognitive abilities of humans," Caramancion added. "We aim to refine our testing protocols, explore new LLMs, and further investigate the dynamics between human cognition and AI technology in the domain of news fact-checking."

More information: Kevin Matthe Caramancion, News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking, *arXiv* (2023). DOI: 10.48550/arxiv.2306.17176