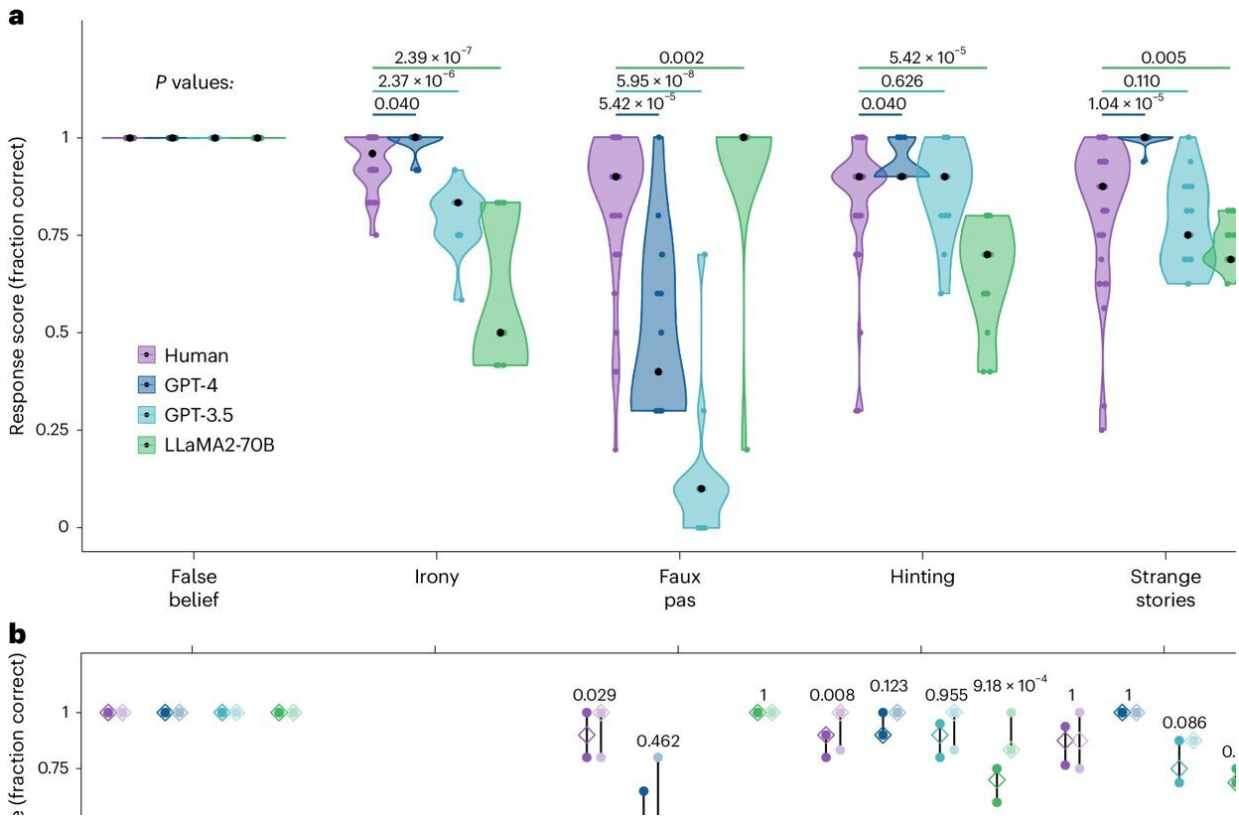
Two types of LLMs found able to equal or outperform humans on theory of mind tests

May 21 2024, by Bob Yirka



Performance of human (purple), GPT-4 (dark blue), GPT-3.5 (light blue) and LLaMA2-70B (green) on the battery of theory of mind tests. a, Original test items for each test showing the distribution of test scores for individual sessions and participants. b, Interquartile ranges of the average scores on the original published items (dark colors) and novel items (pale colors) across each test.

Credit: *Nature Human Behaviour* (2024). DOI: 10.1038/s41562-024-01882-z

An international team of psychologists and neurobiologists has found via experimentation that two types of LLMs are able to equal or outperform humans on theory of mind tests. In their study reported in the journal *Nature Human Behavior*, the group administered theory of mind tests to volunteers and compared the average results with those from two types of LLMs.

Over the past several years, large language models (LLMs) such as ChatGPT have improved to the point that they have now been made available for general use to the public. They have also grown steadily in their abilities. One new ability is to infer mood—hidden meanings or the mental state of a human user.

In this new study, the research team wondered whether the abilities of LLMs have advanced to the point that they can perform theory of mind tasks on par with humans.

Theory of mind tasks were designed by psychologists to measure the mental and/or emotional state of a person during social interactions. Prior research has shown that humans use a variety of cues to signal their mental state to others, with the aim of communicating information without being specific.

Prior research has also shown that humans excel at picking up on such cues, but other animals don't. So many in the field consider it impossible for a computer to pass such tests. The research team tested several LLMs to see how well they would compare to a crowd of humans taking the

same tests.

The researchers analyzed data from 1,907 volunteers who took standard theory of mind tests and compared the results with those of multiple LLMs, such as Llama 2-70b and GPT-4. Both groups answered five types of questions, each designed to measure things like a faux pas, irony or the truth of a statement. Each was also asked to answer "false belief" questions that are often administered to children.

The researchers found that the LLMs quite often equaled the performance of the humans, and sometimes did better. More specifically, they found that GPT-4 was the best of the bunch in five main types of tasks, while Llama-2 scores were much worse than other types of LLMs or humans, in some cases, but was much better at some other types of questions.

According to the researchers, the experiment shows that LLMs are currently able to perform comparably to humans on theory of mind tests, though they are not suggesting that such models are as smart or smarter than humans, or more intuitive in general.

More information: James W. A. Strachan et al, Testing theory of mind in large language models and humans, *Nature Human Behaviour* (2024). DOI: 10.1038/s41562-024-01882-z

Citation: Two types of LLMs found able to equal or outperform humans on theory of mind tests (2024, May 21) retrieved 16 June 2024 from https://techxplore.com/news/2024-05-llms-equal-outperform-humans-theory.html