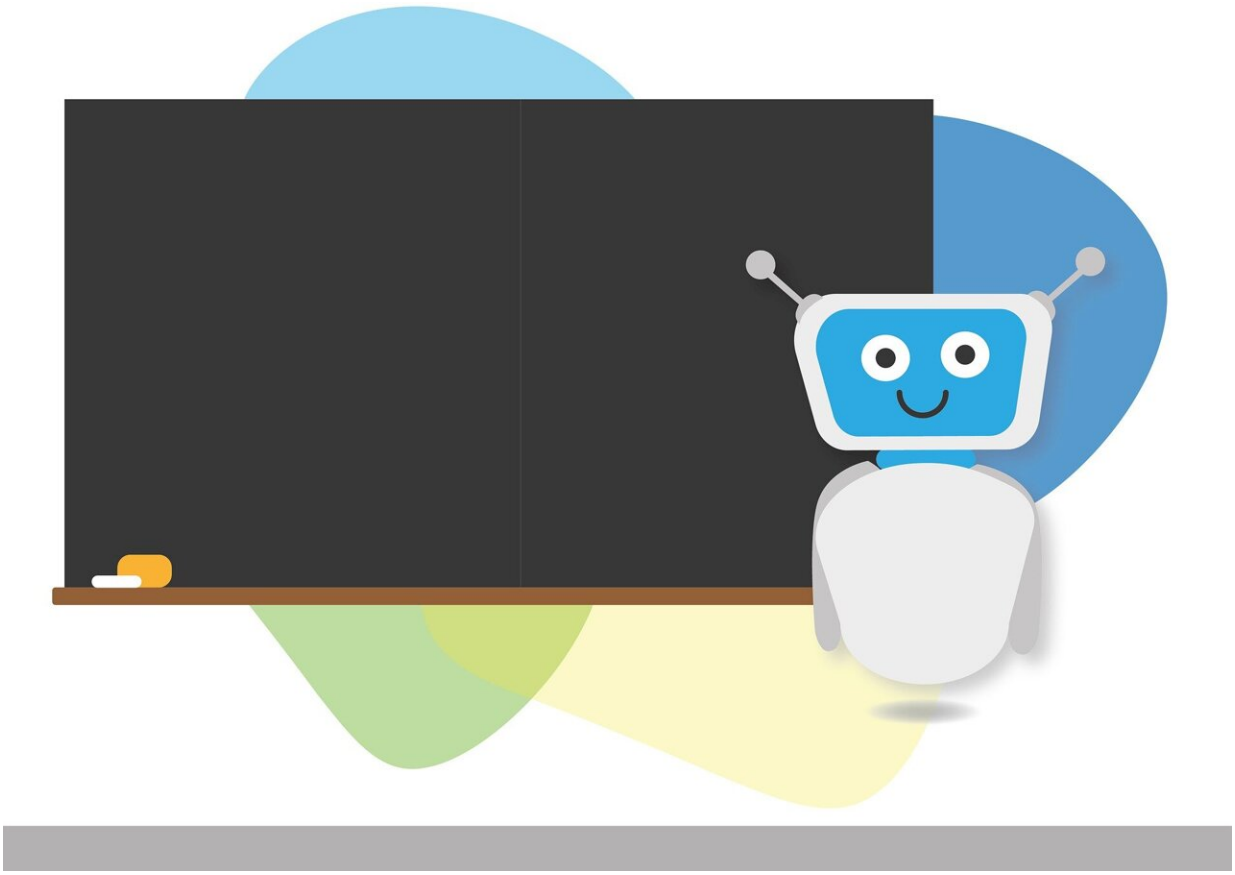
AI poses no existential threat to humanity, new study finds

August 20 2024



Credit: Pixabay/CC0 Public Domain

ChatGPT and other large language models (LLMs) cannot learn independently or acquire new skills, meaning they pose no existential

threat to humanity, according to new research from the University of Bath and the Technical University of Darmstadt in Germany.

The study, published today as part of the proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)—the premier international conference in natural language processing—reveals that LLMs have a superficial ability to follow instructions and excel at proficiency in language, however, they have no potential to master new skills without explicit instruction. This means they remain inherently controllable, predictable and safe.

The research team concluded that LLMs, which are being trained on ever larger datasets, can continue to be deployed without safety concerns, though the technology can still be misused.

With growth, these models are likely to generate more sophisticated language and become better at following explicit and detailed prompts, but they are highly unlikely to gain complex reasoning skills.

"The prevailing narrative that this type of AI is a threat to humanity prevents the widespread adoption and development of these technologies, and also diverts attention from the genuine issues that require our focus," said Dr. Harish Tayyar Madabushi, computer scientist at the University of Bath and co-author of the new study on the "emergent abilities" of LLMs.

The collaborative research team, led by Professor Iryna Gurevych at the Technical University of Darmstadt in Germany, ran experiments to test the ability of LLMs to complete tasks that models have never come across before—the so-called emergent abilities.

As an illustration, LLMs can answer questions about social situations without ever having been explicitly trained or programmed to do so.

While previous research suggested this was a product of models "knowing" about social situations, the researchers showed that it was in fact the result of models using a well-known ability of LLMs to complete tasks based on a few examples presented to them, known as "in-context learning" (ICL).

Through thousands of experiments, the team demonstrated that a combination of LLMs ability to follow instructions (ICL), memory and linguistic proficiency can account for both the capabilities and limitations exhibited by LLMs.

Dr. Tayyar Madabushi said, "The fear has been that as models get bigger and bigger, they will be able to solve new problems that we cannot currently predict, which poses the threat that these larger models might acquire hazardous abilities including reasoning and planning.

"This has triggered a lot of discussion—for instance, at the AI Safety Summit last year at Bletchley Park, for which we were asked for comment—but our study shows that the fear that a model will go away and do something completely unexpected, innovative and potentially dangerous is not valid.

"Concerns over the existential threat posed by LLMs are not restricted to non-experts and have been expressed by some of the top AI researchers across the world."

However, Dr. Madabushi maintains this fear is unfounded as the researchers' tests clearly demonstrated the absence of emergent complex reasoning abilities in LLMs.

"While it's important to address the existing potential for the misuse of AI, such as the creation of fake news and the heightened risk of fraud, it would be premature to enact regulations based on perceived existential

threats," he said.

"Importantly, what this means for end users is that relying on LLMs to interpret and perform complex tasks which require complex reasoning without explicit instruction is likely to be a mistake. Instead, users are likely to benefit from explicitly specifying what they require models to do and providing examples where possible for all but the simplest of tasks."

Professor Gurevych added, "Our results do not mean that AI is not a threat at all. Rather, we show that the purported emergence of complex thinking skills associated with specific threats is not supported by evidence and that we can control the learning process of LLMs very well after all. Future research should therefore focus on other risks posed by the models, such as their potential to be used to generate fake news."

More information: Are Emergent Abilities in Large Language Models just In-Context Learning? aclanthology.org/2024.acl-long.279/

Provided by University of Bath

Citation: AI poses no existential threat to humanity, new study finds (2024, August 20) retrieved 21 August 2024 from https://techxplore.com/news/2024-08-ai-poses-existential-threat-humanity.html