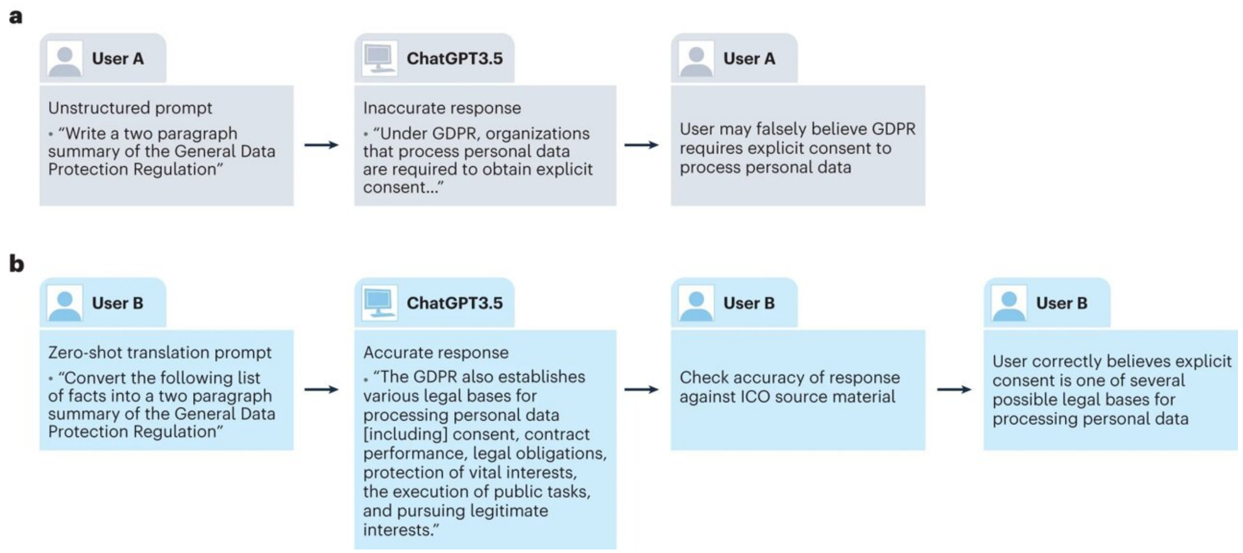
TechXplore

Large language models pose risk to science with false answers, says study

November 20 2023



Unstructured prompt
- "Write a two paragraph summary of the General Data Protection Regulation"

Inaccurate response
- "Under GDPR, organizations that process personal data are required to obtain explicit consent..."

User may falsely believe GDPR requires explicit consent to process personal data

Zero-shot translation prompt
- "Convert the following list of facts into a two paragraph summary of the General Data Protection Regulation"

Accurate response
- "The GDPR also establishes various legal bases for processing personal data [including] consent, contract performance, legal obligations, protection of vital interests, the execution of public tasks, and pursuing legitimate interests."

Check accuracy of response against ICO source material

User correctly believes explicit consent is one of several possible legal bases for processing personal data

Two hypothetical use cases for LLMs based on real prompts and responses demonstrate the effect of inaccurate responses on user beliefs. Credit: *Nature Human Behaviour* (2023). DOI:10.1038/s41562-023-01744-0

Large Language Models (LLMs) pose a direct threat to science because of so-called "hallucinations" (untruthful responses), and should be restricted to protect scientific truth, says a new paper from leading Artificial Intelligence researchers at the Oxford Internet Institute.

The paper by Professors Brent Mittelstadt, Chris Russell, and Sandra Wachter has been published in *Nature Human Behaviour*. It explains, "LLMs are designed to produce helpful and convincing responses without any overriding guarantees regarding their accuracy or alignment with fact."

One reason for this is the data the technology uses to answer questions does not always come from a factually correct source. LLMs are trained on large datasets of text, usually taken from online sources. These can contain false statements, opinions, and creative writing among other types of non-factual information.

Professor Mittelstadt explains, "People using LLMs often anthropomorphize the technology, where they trust it as a human-like information source. This is, in part, due to the design of LLMs as helpful, human-sounding agents that converse with users and answer seemingly any question with confident-sounding, well-written text. The result of this is that users can easily be convinced that responses are accurate even when they have no basis in fact or present a biased or partial version of the truth."

To protect science and education from the spread of bad and biased information, the authors argue, clear expectations should be set around what LLMs can responsibly and helpfully contribute. According to the paper, "For tasks where the truth matters, we encourage users to write translation prompts that include vetted, factual Information."

Professor Wachter says, "The way in which LLMs are used matters. In the scientific community, it is vital that we have confidence in factual information, so it is important to use LLMs responsibly. If LLMs are used to generate and disseminate scientific articles, serious harms could result."

Professor Russell adds, "It's important to take a step back from the opportunities LLMs offer and consider whether we want to give those opportunities to a technology just because we can."

LLMs are currently treated as knowledge bases and used to generate information in response to questions. This makes the user vulnerable both to regurgitated false information that was present in the training data and to "hallucinations"—false information spontaneously generated by the LLM that was not present in the training data.

To overcome this, the authors argue, LLMs should instead be used as "zero-shot translators." Rather than relying on the LLM as a source of relevant information, the user should simply provide the LLM with appropriate information and ask it to transform it into a desired output. For example, rewriting bullet points as a conclusion or generating code to transform scientific data into a graph.

Using LLMs in this way makes it easier to check that the output is factually correct and consistent with the provided input.

The authors acknowledge that the technology will undoubtedly assist with scientific workflows but are clear that scrutiny of its outputs is key to protecting robust science.

"To protect science we must use LLMs as zero-shot translators," lead author Director of Research, Associate Professor and Senior Research Fellow, Dr. Brent Mittelstadt, Oxford Internet Institute.

More information: Mittelstadt, B. et al, To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour* (2023). DOI: 10.1038/s41562-023-01744-0. www.nature.com/articles/s41562-023-01744-0

Provided by University of Oxford

Citation: Large language models pose risk to science with false answers, says study (2023, November 20) retrieved 28 April 2024 from https://techxplore.com/news/2023-11-large-language-pose-science-false.html