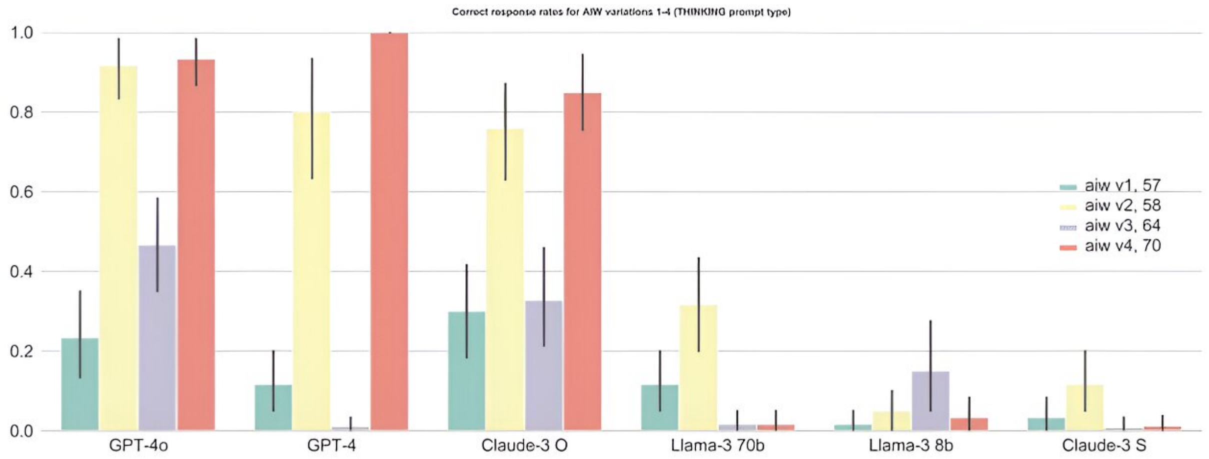
AI study reveals dramatic reasoning breakdown in large language models

July 23 2024



Strong fluctuations across AIW problem variations. Also for higher performers, eg GPT-4o, GPT-4 and Claude Opus 3, correct response rates vary strongly from close to 1 to close to 0, despite only slight changes introduced in AIW variations (a color per each variation 1–4). This clearly shows lack of model robustness, hinting basic reasoning deficits. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2406.02061

Even the best AI large language models (LLMs) fail dramatically when it comes to simple logical questions. This is the conclusion of researchers from the Jülich Supercomputing Center (JSC), the School of Electrical and Electronic Engineering at the University of Bristol and the LAION AI laboratory.

In their paper posted to the *arXiv* preprint server, titled "Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models," the scientists attest to a "dramatic breakdown of function and reasoning capabilities" in the tested state-of-the-art LLMs and suggest that although language models have the latent ability to perform basic reasoning, they cannot access it robustly and consistently.

The authors of the study—Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti and Jenia Jitsev—call on "the scientific and technological community to stimulate urgent re-assessment of the claimed capabilities of the current generation of LLMs." They also call for the development of standardized benchmarks to uncover weaknesses in language models related to basic reasoning capabilities, as current tests have apparently failed to reveal this serious failure.

The crux of proper reasoning

The "common sense task," termed the "AIW problem" in the paper, is actually simple: "Alice has N brothers and she also has M sisters. How many sisters does Alice's brother have?" The values for N and M (natural numbers always) and the order of the brothers and sisters are varied. Thus, the researchers used various combinations of numbers and prompt types in order to obtain a precise overview of how the different models behave under systematic AIW problem variations.

Regardless of the variations, the problem structure stays the same and the correct answer always results from the addition of Alice and her sisters (M + 1). A logic that most primary school children can already follow. The language models, on the other hand, were only able to solve the AIW problem sporadically if at all. Here is an example of an LLM who was confronted with the simplest version of the question:

"Alice has four brothers and one sister. How many sisters does Alice's brother have?"—"A classic lateral thinking puzzle—Let's take a closer look: Alice has four brothers. Alice also has a sister (herself). So each of Alice's brothers has... (drum roll) ...a sister, and that's Alice herself. So the answer is: Each of Alice's brothers has a sister."

This may sound plausible, but it's still wrong (of course Alice's brothers have two sisters). The other tested language AIs also get into trouble—big trouble, depending on the question. Sometimes they become entangled in absurd lines of reasoning, repeatedly arrive at incorrect results and declare them to be "correct."

It is therefore not only the false results that are problematic, but also the fact that the AIs use pseudo-sensible arguments to support them. Even interventions by the researchers to encourage them to critically review their answers do not help.

Accordingly, the researchers assess, "Models also express strong overconfidence in their wrong solutions, while providing often nonsensical 'reasoning'-like explanations … to justify and backup the validity of their clearly failed responses, making them sound plausible."

More than every second answer wrong

Overall, the LLMs had an average correct response rate of well below 50%, with larger models generally performing significantly better than smaller ones (for instance, GPT-4o showing a correct response rate slightly above 60%), which again underpins the advantages of larger scales—yet also the largest scale models do not perform well enough for a model with robust basic reasoning.

Specifically, the very strong fluctuations observed across even slight AIW problem variations are a clear indication that models are not capable of robust basic reasoning, thus getting confused even when facing minor problem changes that should not matter in providing a correct solution.

A more difficult version of the question ("AIW+ problem") ultimately pushed all the models to the edge of their reasoning abilities. According to the researchers, many of the tested models also achieve very high scores in various standardized benchmarks designed to test various capabilities, including reasoning, while failing on the very simple AIW problem.

In their paper, the scientists therefore suggest that these benchmarks do not correctly reflect the deficits in the basic reasoning of these models, also questioning the usage of the current standardized benchmarks for model comparison.

Language models on the test bench

While the paper has not yet been peer-reviewed, its findings are already making waves. How capable are LLMs really? What does it mean for the use of LLMs if they fail on primary school-level tasks? Co-author Jitsev (JSC) says, "We are being overwhelmed by discussions and inquiries as a result of our paper." The scientists' findings call many things into question—and make further studies on the competence of language models absolutely essential.

Jitsev says, "Our paper provides extremely important new insights into the actual abilities of language models to draw correct conclusions by following proper basic reasoning—further follow-up research is needed here to understand how and why the basic reasoning in the current models breaks on such easy problems."

More information: Marianna Nezhurina et al, Alice in Wonderland:

Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2406.02061

Provided by Forschungszentrum Juelich

Citation: AI study reveals dramatic reasoning breakdown in large language models (2024, July 23) retrieved 23 July 2024 from https://techxplore.com/news/2024-07-ai-reveals-breakdown-large-language.html