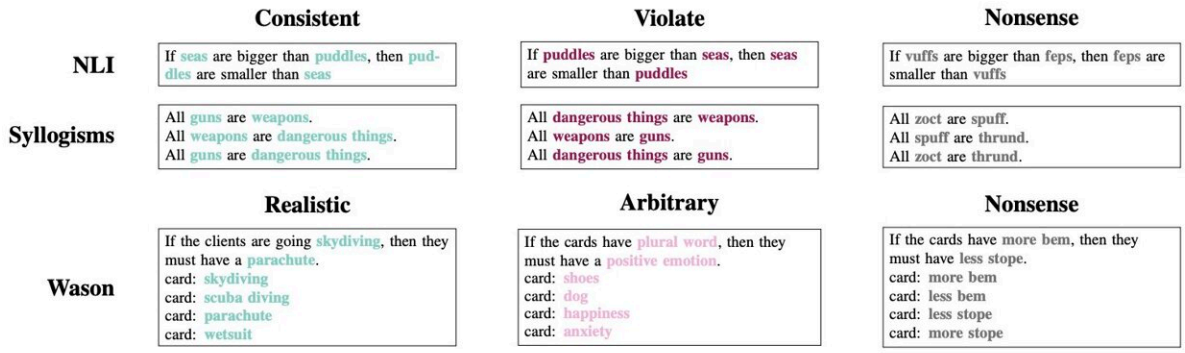
Large language models make human-like reasoning mistakes, researchers find

July 16 2024



Manipulating content within fixed logical structures. In each of the author's three datasets, they instantiate different versions of the logical problems. Different versions of a problem offer the same logical structures and tasks but instantiated with different entities or relationships between those entities. The relationships in a task may either be consistent with, or violate real-world semantic relationships, or may be nonsense, without semantic content. In general, humans and models reason more accurately about belief-consistent or realistic situations or rules than belief-violating or arbitrary ones. Credit: Lampinen et al

Large language models (LLMs) can complete abstract reasoning tasks, but they are susceptible to many of the same types of mistakes made by humans. Andrew Lampinen, Ishita Dasgupta, and colleagues tested state-of-the-art LLMs and humans on three kinds of reasoning tasks: natural language inference, judging the logical validity of syllogisms, and the

Wason selection task.

The findings are published in *PNAS Nexus*.

The authors found the LLMs to be prone to similar content effects as humans. Both humans and LLMs are more likely to mistakenly label an invalid argument as valid when the semantic content is sensical and believable.

LLMs are also just as bad as humans at the Wason selection task, in which the participant is presented with four cards with letters or numbers written on them (e.g., "D," "F," "3," and "7") and asked which cards they would need to flip over to verify the accuracy of a rule such as "if a card has a 'D' on one side, then it has a '3' on the other side."

Humans often opt to flip over cards that do not offer any information about the validity of the rule but that test the contrapositive rule. In this example, humans would tend to choose the card labeled "3," even though the rule does not imply that a card with "3" would have "D" on the reverse. LLMs make this and other errors but show a similar overall error rate to humans.

Human and LLM performance on the Wason selection task improves if the rules about arbitrary letters and numbers are replaced with socially relevant relationships, such as people's ages and whether a person is drinking alcohol or soda. According to the authors, LLMs trained on human data seem to exhibit some human foibles in terms of reasoning—and, like humans, may require formal training to improve their logical reasoning performance.

More information: Language models, like humans, show content effects on reasoning tasks, *PNAS Nexus* (2024). DOI: 10.1093/pnasnexus/pgae233. academic.oup.com/pnasnexus/art …

/3/7/pgae233/7712372

Provided by PNAS Nexus