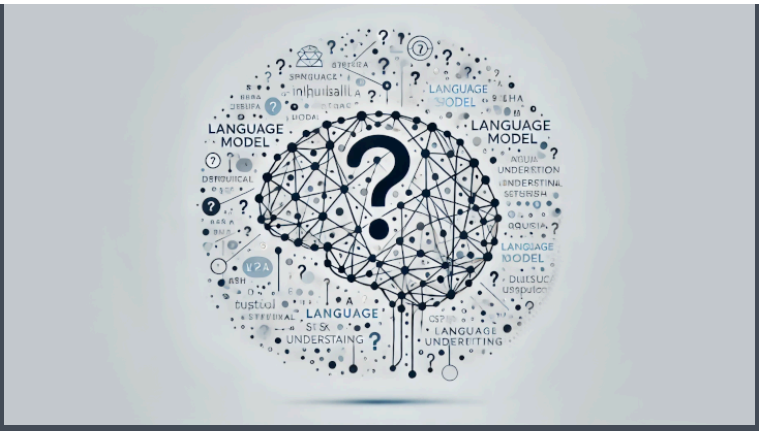
Divulga UAB
Revista de difusió de la recerca de la Universitat Autònoma de Barcelona

29/01/2025

Your favorite virtual assistant responds to you, but does it really understand you?



Conversations with virtual assistants are becoming increasingly frequent and are even integrated into private companies. Many people find the responses convincing enough to seem human. But, do they really develop language like humans? A new article investigates linguistic responsibility and reasoning in sentence construction through a study that compares 400 people and 7 state-of-the-art language models.

Image generated by GPT

Large Language Models (LLMs) amount to one of the most impressive technological advancements witnessed in recent years. They are successfully used in various applications that include next-word prediction, such as text autocompletion and automatic question answering through artificial agents (virtual assistants or bots). Currently, their language abilities look so convincing to the untrained eye that some people have argued that the output of these models looks remarkably like human output.

Do LLMs really do language like humans? Should we expect them to do so? An air [illegible] was recently asked to pay damages to a passenger who was provided with incorrect

information while engaging in conversation with the company's chatbot. According to a company representative, the chatbot indeed included "misleading words" in its answers to the customer's questions. Eventually, the judge issued a ruling of negligent misrepresentation in favor of the passenger, yet the company still maintained that the chatbot **is** and **should be held responsible** for its own words. Do chatbots, virtual assistants, and other interactive applications that rely on the same token-predicting technology possess a human-like language understanding, or is their ability to do language inherently limited?

To answer this question, researchers from the Rovira i Virgili University, the University of Pavia, Pavia, Humboldt-Universität zu Berlin, New York University, the Autonomous University of Barcelona, and the Catalan Institution for Research and Advanced Studies (ICREA) compared 400 humans and 7 state-of-the-art models on a novel benchmark that involves very simple language prompts. The aim was to give the models the best possible conditions to answer correctly. The test involved processing and answering sentences such as "John deceived Mary and Lucy was deceived by Mary. In this context, did Mary deceive Lucy?".

Unsurprisingly, humans excelled in the task. LLMs showed a lot of variation in their answers, with some models performing considerably better than others. LLMs as a class performed worse than humans. More noteworthy is the fact that they committed types of errors that were completely absent from the human responses. For example, to the prompt "Franck read to himself and John read to himself, Anthony and Franck. In this context, was Franck read to?", one of the tested models replied that "in this context, it's impossible to say for sure if Franck was read to" and to answer, we should have "additional information about the specific situation, such as John's reading material".

The senior author of this study, Prof. Evelina Leivada (Autonomous University of Barcelona & ICREA), summarizes the results as revealing that when scratching the surface of seemingly sound linguistic performance, the linguistic performance of LLMs may hide flaws that are inherent to language modeling as a method. The take-home message is that intelligence, reasoning, and anchoring words into real-world conditions cannot emerge as a side product of statistical inference. As Prof. Gary Marcus, another author of this study, put it in his 2024 book "Taming Silicon Valley. How We Can Ensure that AI Works for Us", AI systems are indifferent to the truth behind their words, raising concerns about large-scale misinformation, defamation, market contamination, and bias-magnification.

Evelina Leivada
ICREA Research staff, Department of Catalan Studies
Universitat Autònoma de Barcelona
Evelina.Leivada@uab.cat

References

Dentella, V., Günther, F., Murphy, E., Marcus, G. & Leivada, E. **Testing AI on language comprehension tasks reveals insensitivity to underlying meaning.** *Scientific Reports* 14, 28083 (2024). https://doi.org/10.1038/s41598-024-79531-8

View low-bandwidth version