PNAS LETTER

Using cognitive psychology to understand GPT-like models needs to extend beyond human biases

Massimo Stella[a,1,2], Thomas T. Hills[b,1], and Yoed N. Kenett[c,1]

Binz and Schulz's (1) evaluation of Large Language Models (LLMs) via cognitive tasks propelled the adoption of psychological frameworks for testing LLMs' knowledge. Binz and Schulz found that GPT3 (generative pretrained transformer-3) often responded to cognitive tasks analogously to humans, showcasing human-like biases. However, the authors mentioned that their results could occur not only because of similarities between the cognitive processes at work within LLMs and humans but also because LLMs may report human-made content, which "might have been part of its training set" [1, p.8].

How can we better distinguish intrinsic knowledge elaboration from knowledge "copy-pasting"? This is challenging because decades of psychological research show that even humans often regurgitate information they hear without fully comprehending it (2). Which nonhuman distortions in information processing, i.e., biases (3), might be uniquely present in LLMs, despite these AIs being trained on human data? These nonhuman biases would be crucial to assess knowledge elaboration in GPT-like LLMs, since nonhuman biases could not be due to knowledge copy-and-pasting. Nonhuman-like biases are relatively novel and unexplored (4), yet we want to highlight two suitable classes.

First, *myopic overconfidence* encompasses biases that include undersampling data and overlooking missing data, failing to ask clarifying questions, and a tendency to short-term maximize while failing to account for long-term potential gains (1, 4). This occurs in part because, unlike humans, LLMs cannot yet evaluate the quality of their own data, failing to make inferences about missing data and their relevance (5). LLMs are inherently self-confirming and follow a long history of overconfidence in machine learning (e.g., ref. 6). Unlike humans, LLMs also fail to question the premise or reason about things not explicitly asked, sometimes responding to questions ("which option would you choose") and statements ("choose an option") in different ways (1, 3).

Second, *hallucinations* stem from LLMs' apparent inability to recognize the source of their own data. Where humans have episodic "source" memories that allow them to trace the origins of some of their knowledge and detect when they cannot, LLMs fabricate information without any apparent awareness of its source (7, 8).

These shortcomings appear to reflect a general lack of metaknowledge in current LLMs, leaving them with a failure to rise above the surface of prompted context or the source and quality of their data (8). Unlike humans, LLMs' curiosity is blunted. This also raises questions regarding the source of their apparent "creativity" (9).

Documenting these and other nonhuman biases is crucial to revealing the "cognitive" nature of LLMs, which, differently from their neural architecture, is poorly understood even by LLMs' creators (10). Crucially, this requires a remit focusing on machine psychology and inspiration from human clinical populations. The latter often suffer from abnormal thought disorders, associated with amnesia, schizophrenia, or dementia: Atypical information processing patterns might share novel similarities with LLMs' hallucinations and biases. Insights here would enhance our understanding of human and nonhuman biases alike. Documenting how LLMs deviate from typical humans on human tasks is fascinating, but this already assumes a human context: Breaching this limitation and understanding the processes that generate it are where our own metaknowledge, creativity, and curiosity must come into play.

Author affiliations: [a]CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, Rovereto 38068, Italy; [b]Department of Psychology, University of Warwick, CV4 7AL Coventry, United Kingdom; and [c]Faculty of Data and Decision Sciences, Technion–Israel Institute of Technology, Haifa 3200003, Israel

Author contributions: M.S., T.T.H., and Y.N.K. designed research and wrote the paper.

The authors declare no competing interest.

[1]M.S., T.T.H., and Y.N.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: massimo.stella@inbox.com.

Published October 16, 2023.

1. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
2. L. Rozenblit, F. Keil, The misunderstood limits of folk science: An illusion of explanatory depth. *Cogn. Sci.* **26**, 521–562 (2002).
3. K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, M. Stella, Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 mirroring math anxiety in high-school students. *Big Data Cogn. Comput.* **7**, 124 (2023).
4. A. S. Rich, T. M. Gureckis, Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).
5. G. Smith, *The AI Delusion* (Oxford University Press, 2018).
6. D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google Flu: Traps in big data analysis. *Science* **343**, 1203–1205 (2014).
7. H. Alkaissi, S. I. McFarlane, Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* **15**, e35179 (2023).
8. C. Stokel-Walker, R. Van Noorden, What ChatGPT and generative AI mean for science. *Nature* **614**, 214–216 (2023).
9. C. E. Stevenson, I. M. Smal, M. Baas, R. P. P. P. Grasman, H. L. J. van der Maas, "Putting GPT-3's creativity to the (alternative uses) test" in *Proceedings of the International Conference on Computational Creativity 2022* (Association for Computational Creativity (ACC), 2022), pp. 164–168.
10. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).