PNAS COMMENTARY

Counterfeit judgments in large language models

Matjaž Perc[a,b,c,d,1]

The idea of creating machines that could imitate or reproduce human thought has deep historical roots, long predating the digital computer. In Greek mythology, Hephaestus forged mechanical servants of gold that could think and speak, while in Jewish folklore, the Golem was an artificial being that could be animated by Hebrew letters and mystical incantations. By the Middle Ages and Renaissance, thinkers such as Ramon Llull and later Gottfried Wilhelm Leibniz began to imagine logical machines—devices that might reason mechanically using symbolic systems. Leibniz's vision of a calculus ratiocinator was an especially prescient forerunner of formal logic and, ultimately, computer science (1). The Enlightenment's mechanical philosophers, from Descartes to La Mettrie, further speculated on the human mind as a machine—a system of parts that might one day be replicated artificially. In the mid-20th century, when advances in formal logic, mathematics, and electronics finally converged, Alan Turing proposed an operational definition of intelligence through the imitation game, now known as the Turing Test (2). This was a turning point in that, rather than asking what thinking is, Turing asked whether machine behavior could be indistinguishable from human behavior. The field of artificial intelligence was formally established thereafter and the ambition that machines might replicate human judgment was transformed from philosophical speculation into an empirical and technological enterprise of today (3). However, the mechanisms by which artificial intelligence arrives at judgments—and how these differ from human reasoning—remain poorly understood. The study by Loru et al. (4), in the current issue of PNAS, provides a systematic investigation of how large language models (LLMs) operationalize key evaluative concepts such as reliability and bias, and what this reveals about the broader epistemic transformations introduced by automated judgment.

Namely, the latest versions of LLMs, such as ChatGPT, Deepseek, Gemini, Llama, and Mistral all pass the Turing Test, and often more commonly so than humans. Not surprisingly, research has already claimed that LLMs may serve as useful surrogates for humans in social science research (5), market research (6), as well as healthcare (7, 8). LLMs also enjoy widespread use in everyday tasks, such as text generation, language translation, summarization, and support in education (9). Moreover, arguments have actually been put forth in favor of LLMs over humans, which range from cost-effectiveness and scalability to the promise of troves of synthetic data where it would otherwise be scarce (10, 11). But as these models are rapidly becoming integral to how societies process and evaluate information, and as they increasingly participate in evaluative tasks that were traditionally reserved for humans, the alarm bells are ringing ever louder (12, 13). How is judgment itself operationalized in LLMs?—A fundamental question that Loru et al. (4) now answer.

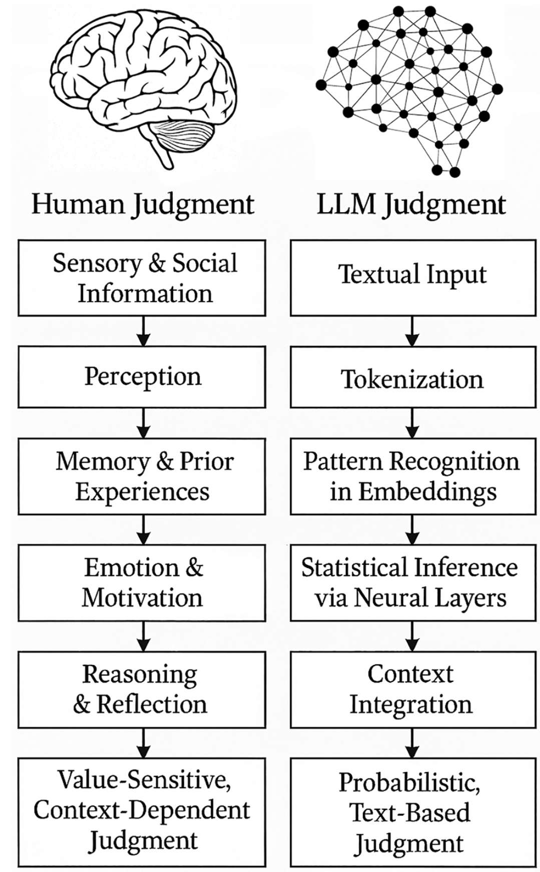

Fig. 1. Parallel information pathways in human and artificial judgment. On the *Right* hand side, we have a typical information pathway of a human to reach a certain judgment. On the *Left* hand side, we have a generic large language model pathway with steps that very roughly correspond to human action. Research done by Loru et al. (4) finds consistent differences in judgments arrived at one way or the other. In fact, they conclude that lexical associations and statistical priors influence evaluations in ways that differ from human reasoning and normative criteria. This divergence is termed *epistemia*—the illusion of knowledge that emerges when surface plausibility replaces verification. More to the point, these findings are indicative not only of systematic performance asymmetries, but they reveal also a deep shift in the heuristics that underlie evaluative processes and raise fundamental questions about delegating judgment to LLMs.

Author affiliations: [a]Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor 2000, Slovenia; [b]Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor 2000, Slovenia; [c]Department of Physics, Kyung Hee University, Dongdaemun-gu, Seoul 02447, Republic of Korea; and [d]University College, Korea University, Seongbuk-gu, Seoul 02841, Republic of Korea

Author contributions: M.P. wrote the paper.

The author declares no competing interest.

See companion article, "The simulation of judgment in LLMs," 10.1073/pnas.2518443122.

[1]Email: matjaz.perc@gmail.com.

Published November 24, 2025.

To that effect the authors benchmarked six LLMs against expert evaluations from NewsGuard and Media Bias/Fact Check as well as against human participants engaged in a controlled experimental task. To ensure comparability, they implemented a structured agentic framework: Both models and human subjects followed the same stepwise evaluation process—selecting assessment criteria, retrieving relevant information, and providing written justifications. Importantly, the focus was not on the news content itself, but on how evaluative reasoning unfolds within and across agents.

At the surface level, LLMs often appeared to align closely with expert classifications, especially in flagging unreliable or low-credibility outlets. However, this apparent convergence masked fundamental differences in how the judgments were produced. Human participants drew on contextual understanding, content comprehension, and pragmatic reasoning to assess credibility (14). LLMs, by contrast, relied heavily on lexical associations, statistical priors, and surface-level linguistic cues embedded in their training data. The models thus operated through pattern-based approximation rather than deliberative or normative reasoning, thus motivating the term "counterfeit judgments." These differences are summarized in Fig. 1.

The study by Loru et al. (4), in the current issue of PNAS, provides a systematic investigation of how large language models (LLMs) operationalize key evaluative concepts such as reliability and bias, and what this reveals about the broader epistemic transformations introduced by automated judgment.

This reliance on form over substance generated a distinctive dynamic that the authors term epistemia—the illusion of knowledge that emerges when linguistic plausibility substitutes for verification (4). In this mode, coherent and well-structured language can appear truthful regardless of factual grounding. Such epistemia poses epistemic risks when automated systems are used to evaluate information in environments already characterized by misinformation and cognitive overload (15, 16).

The study also uncovered systematic asymmetries in LLM evaluations. Right-leaning outlets were more likely to be classified as unreliable, reflecting how the models' training distributions and linguistic priors may encode and reproduce sociopolitical biases. In addition, models tended to exhibit what the authors describe as a "skepticism bias"—a tendency to overreject accurate information—mirroring patterns previously observed in human psychology (17). Yet, despite these superficial similarities, the underlying mechanisms diverged sharply: LLMs mimic certain behavioral regularities of human judgment without replicating the cognitive processes that generate them.

When examining the evaluative criteria, both LLMs and human participants rated "factual accuracy" as the most important dimension. However, their interpretations of accuracy differed profoundly. Human evaluators tended to assess accuracy through comprehension and contextual reasoning, while LLMs inferred it statistically from textual regularities and correlations learned during training. Secondary criteria further highlighted this divide: LLMs emphasized "ownership transparency" and institutional markers consistent with fact-checking protocols, while humans focused on rhetorical features such as tone, clarity, and emotional neutrality. These tendencies echo cognitive heuristics like processing fluency, whereby ease of reading enhances perceived truthfulness—but in LLMs, such effects emerge through structural rather than psychological mechanisms.

The implications of the performed research extend far beyond news classification. As LLMs are increasingly embedded into decision-making infrastructures—from social media moderation and information filtering to policy analysis, law, and scientific evaluation—we warn that society may be witnessing a structural shift in how evaluation itself is defined and operationalized. Automated systems transform evaluative reasoning from a deliberative, context-aware process into one driven by pattern recognition and probabilistic approximation. This shift introduces new forms of epistemic opacity, where judgments appear authoritative but are underpinned by mechanisms opaque to human scrutiny.

We also argue that such a transformation marks a new phase in the evolution of the information ecosystem. Whereas early research on social media focused on the problems of volume and velocity—the sheer overload of information (18)—the current challenge concerns the nature of judgment itself. When plausibility replaces verification and pattern matching substitutes for reasoning, the architecture of collective knowledge becomes susceptible to distortions that arise not from misinformation alone (15, 16), but from the epistemic procedures used to assess it.

To mitigate these risks, we call, together with many other researchers (4, 12, 13, 19), for greater transparency in model reasoning, sustained human oversight, and the development of hybrid approaches that integrate statistical models with retrieval-based evidence and explicit reasoning and moral criteria (20). Such frameworks could help disentangle factual reliability from ideological or stylistic cues, allowing LLMs to support rather than supplant human judgment.

Last, we wish to underscore an ever more apparent paradox of the AI era: LLMs can replicate the appearance of human evaluation without reproducing its substance. Their judgments may seem accurate, but they are built upon fundamentally different epistemic foundations. Recognizing and addressing this divergence is essential if automated systems are to enhance—rather than erode—the integrity of human knowledge and deliberation.

ACKNOWLEDGMENTS. M.P. was supported by the Slovenian Research and Innovation Agency (Grant No. P1-0403).

1. W. Lenzen, "Leibniz and the calculus ratiocinator" in *Technology and Mathematics: Philosophical and Historical Investigations* , S. O. Hansson, Ed. (Springer, 2018), pp. 47–78.
2. A. M. Turing, "Computing machinery and intelligence" in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* , R. Epstein, G. Roberts, G. Beber, Eds. (Springer, 2009), pp. 23–65.
3. W. X. Zhao *et al*., A survey of large language models. arXiv [Preprint] (2025). http://arxiv.org/abs/2303.18223 (Accessed 12 October 2025).
4. E. Loru *et al*., The simulation of judgment in LLMs. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2518443122 (2025).
5. I. Grossmann *et al*., AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
6. J. Brand, A. Israeli, D. Ngwe, "Using LLMs for market research." SSRN (2023). https://ssrn.com/abstract=4395751.
7. M. Cascella, J. Montomoli, V. Bellini, E. Bignami, Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J. Med. Syst.* **47**, 33 (2023).
8. S. Uygun İlikhan, M. Özer, M. Perc, H. Tanberkan, Y. Ayhan, Complementary use of artificial intelligence in healthcare. *Med. J. West. Black Sea* **9**, 7–17 (2025).
9. M. Perc, M. Özer, Disappearing minds in the age of artificial intelligence. *Insan Toplum* **15**, 1–9 (2025).
10. J. J. Horton, *Large Language Models as Simulated Economic Agents: What Can we Learn from Homo Silicus?* (National Bureau of Economic Research, 2023).
11. N. Arora, I. Chakraborty, Y. Nishimura, AI-human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. *J. Mark.* **89**, 43–70 (2025).
12. D. C. Dennett, *The Problem with Counterfeit People* (The Atlantic, 2023).
13. Y. Gao, D. Lee, G. Burtch, S. Fazelpour, Take caution in using LLMs as human surrogates. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2501660122 (2025).
14. S. Fiske, *Social Cognition: Selected Works of Susan Fiske* (Routledge, Abingdon, U.K., 2018).
15. M. Del Vicario *et al*., The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016).
16. C. Budak, B. Nyhan, D. M. Rothschild, E. Thorson, D. J. Watts, Misunderstanding the harms of online misinformation. *Nature* **630**, 45–53 (2024).
17. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
18. D. Bawden, L. Robinson, The dark side of information: Overload, anxiety and other paradoxes and pathologies. *J. Inf. Sci.* **35**, 180–191 (2009).
19. M. Binz *et al*., How should the advancement of large language models affect the practice of science?. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121 (2025).
20. V. Capraro, M. Perc, Mathematical foundations of moral preferences. *J. R. Soc. Interface* **18**, 20200880 (2021).