frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 27 May 2025
DOI 10.3389/frai.2025.1592399

OPEN ACCESS

EDITED BY
Amir Zadeh,
Wright State University, United States

REVIEWED BY
J. D. Opdyke,
Sachs Capital Group Asset Management, LLC, United States

*CORRESPONDENCE
Andrea Polonioli
✉ apolonioli@coveo.com

RECEIVED 12 March 2025
ACCEPTED 29 April 2025
PUBLISHED 27 May 2025

CITATION
Polonioli A (2025) Moving LLM evaluation forward: lessons from human judgment research.
Front. Artif. Intell. 8:1592399.
doi: 10.3389/frai.2025.1592399

Moving LLM evaluation forward: lessons from human judgment research

Andrea Polonioli*

Coveo, Quebec City, QC, Canada

This paper outlines a path toward more reliable and effective evaluation of Large Language Models (LLMs). It argues that insights from the study of human judgment and decision-making can illuminate current challenges in LLM assessment and help close critical gaps in how models are evaluated. By drawing parallels between human reasoning and model behavior, the paper advocates moving beyond narrow metrics toward more nuanced, ecologically valid frameworks.

KEYWORDS

LLM, generative AI (GenAI), hallucinations, AI in business, human judgment, judgment and decision making, heuristics & biases

1 Introduction

Large Language Models (LLMs) have become central to the progress of artificial intelligence, powering advances across industries—from healthcare and education to legal analysis and creative writing (Chowdhery et al., 2022; Touvron et al., 2023). The public release of ChatGPT in 2022 marked a turning point, introducing LLMs into everyday discourse and positioning them as general-purpose intelligence systems. Yet despite their impressive versatility, these models often produce surprising errors, raising persistent questions about how to evaluate their reliability and adaptability (Bishop, 2021).

A growing ecosystem of benchmarks has emerged to address this challenge. Factuality assessments such as FELM (Chen et al., 2023), code-focused tasks like HumanEval (Chen et al., 2021), and domain-specific evaluations like SWE-bench Verified (Jimenez et al., 2023) each offer partial insight into model capabilities. Ranking-based platforms like Chatbot Arena (Zheng et al., 2023) have further shaped public perception, rewarding models that perform well in direct comparison. Yet these evaluation strategies remain fragmented and narrow, often incentivizing superficial improvements rather than generalizable progress.

Promising developments within the deep learning community have begun to address these limitations. Notably, Martin et al. (2021) present a framework for evaluating neural networks using structural metrics derived from the models' own weight matrices. Building on theoretical results by Martin and Mahoney (2021), their approach offers a means of assessment that does not rely on external test data. It introduces a different kind of benchmark—one that focuses on the internal properties of a model and the distribution of capacity across its architecture. In doing so, it offers a more nuanced perspective on model quality, extending beyond the fragmented and task-bound metrics that dominate much of today's evaluation landscape. Nevertheless, such contributions have yet to significantly shape the broader discourse, which remains largely driven by surface-level performance and high-profile failure cases (e.g., Silberling, 2024).

Consider this seemingly simple exchange:

> Human: "How many R's are in the word *strawberry*?"
> LLM: "There are two."

> Human: "Actually, there are three—one in the middle and two at the end."
> LLM: "No, count again."

LLMs frequently fail at these kinds of counting tasks, producing confident but incorrect responses. Such errors raise concerns not only about model precision but also about the deeper mechanics of how these systems handle symbolic information and logical sequence processing. Do these failures reflect minor blind spots in token processing, or do they expose more fundamental architectural limitations? Could such mistakes result from asking the wrong kind of question—or using the wrong kind of evaluation? Are these isolated quirks, or signs of broader, generalizable weaknesses? And do different models exhibit systematically different error patterns? Recent evidence suggests yes – different models can have distinctive failure profiles. For example, Martin et al.'s analysis (2020) indicates that model architecture and training influence the types of errors a network is prone to.

Once these questions are raised, it becomes clear that they echo long-standing debates in the study of human judgment. For decades, cognitive scientists have explored how people process information, why they make systematic errors, and whether such errors signal irrationality or adaptive trade-offs. Concepts like bounded rationality (Simon, 1955) and ecological validity (Gigerenzer and Todd, 1999) helped reframe these debates—moving beyond binary success/failure judgments toward more nuanced, context-sensitive models of reasoning. These same ideas, we argue, can enrich the way we approach LLM evaluation.

This paper contends that advancing LLM evaluation requires drawing from the intellectual history of human judgment research. By moving beyond narrow benchmarks and reductive metaphors toward frameworks that foreground trade-offs, context, and structured interventions, we can build a more robust and empirically grounded understanding of what these models can - and cannot - do.

2 Accuracy does not speak with one voice

Just as research in human judgment and decision-making has long been shaped by influential metaphors (e.g., "cognitive illusions" and "biases"), the evaluation of LLMs has similarly gravitated toward evocative language. In particular, "hallucination" has emerged as a dominant descriptor of model error. While some scholars have proposed alternatives like "confabulation," drawn from neuropsychology to describe plausible but incorrect responses in the absence of sufficient information (Smith et al., 2023), others—such as Brender (2023)—have rejected anthropomorphic metaphors altogether, warning that terms like *hallucination* risk projecting human cognitive assumptions onto fundamentally different systems.[1]

The issue with such metaphors is not only that they introduce conceptual baggage or polarize discussion; more critically, they oversimplify the multifaceted nature of model failure. LLM outputs do not merely succeed or fail in binary terms—accuracy manifests across different dimensions. Some errors reflect misalignment with external truth (factuality), while others arise from internal inconsistency, poor calibration, or sensitivity to prompt phrasing.

Hammond's (2007) distinction between coherence and correspondence in human judgment offers a useful lens. Coherence refers to internal consistency—how well a model's outputs logically hang together. This concept is central to the heuristics-and-biases tradition, which often highlights deviations from logical norms (e.g., the conjunction fallacy; Kahneman and Tversky, 1983). Correspondence, by contrast, focuses on alignment with external reality and predictive success, as seen in ecological approaches like fast-and-frugal heuristics (Gigerenzer and Todd, 1999; Polonioli, 2014, 2016). For example, the recognition heuristic can help people make accurate predictions in uncertain environments despite limited information.

Crucially, coherence and correspondence do not always align (Arkes et al., 2016; Katsikopoulos, 2009). Coherence-based evaluations often cast human reasoning in a negative light, while correspondence-based approaches highlight when heuristics yield adaptive, real-world performance. This tension has been instrumental in reshaping how we assess rationality, and it offers a valuable precedent for LLM evaluation. Polonioli (2015) further argues that the coherence–correspondence distinction, while useful, does not exhaust the complexity of cognitive evaluation. Other dimensions—such as context sensitivity and calibration—also matter. As Nisbett and Wilson (1977) famously showed, human judgments are heavily influenced by contextual cues. LLMs exhibit similar fragility: minor prompt variations can yield dramatically different outputs, yet few benchmarks test this.

Despite a growing ecosystem of benchmarks, most focus overwhelmingly on correspondence. Datasets such as FELM (Factuality Evaluation of Large Language Models; Chen et al., 2023) or TruthfulQA (Lin et al., 2022) measure accuracy relative to known facts. These tools are valuable—but they neglect coherence-related errors, such as when models contradict themselves or generate answers that do not align with their own justifications.

Several recent studies hint at the importance of coherence, though not always explicitly. For example, Wang et al. (2022), in Self-Consistency Improves Chain-of-Thought Reasoning in Language Models, show that averaging a model's answers across multiple reasoning paths often improves correctness—suggesting that internal consistency may correlate with better performance. Zhou et al. (2022), in *Least-to-Most Prompting Enables Complex Reasoning*, point out that LLMs sometimes arrive at correct answers via logically invalid chains—indicating that output correctness does not always reflect processing quality.

Other work more directly engages with internal inconsistency. Madaan et al. (2023), in Self-Refine: Iterative Refinement with Self-Feedback, explore prompting models to revise their own outputs—a method that frequently surfaces contradictions and logical analysis failures. Meanwhile, Macmillan-Scott and Musolesi (2024), in Biases and Fallacies in Large Language Models: A Human Reasoning Perspective, test LLMs on known human reasoning biases. Their findings show that while models can replicate certain fallacies, they often do so inconsistently or incoherently—further demonstrating that LLM failure modes do not map cleanly into human patterns.

1 While terms like "reasoning" and "hallucination" are widely used as convenient functional descriptors of model behavior, they should not be taken to imply that LLMs possess cognitive or experiential capacities akin to those of human minds. For a critique of such anthropomorphic metaphors—and of the conceptual risks involved in borrowing language between AI and brain sciences—see Floridi and Nobre (2024).

Despite these developments, there is still no large-scale benchmark dedicated to assessing coherence in LLMs. This is a critical gap. If coherence is key to evaluating the quality of *how* models arrive at answers, then the absence of such a benchmark skews our understanding of model behavior and limits opportunities for targeted improvement. A coherence benchmark would bring at least three benefits:

1 Clarify the coherence–correspondence relationship: It would help disentangle cases where models generate correct answers for the wrong reasons—or coherent but incorrect responses.
2 Test generalization more meaningfully: Stable, consistent reasoning is likely to be more robust across prompt variations and domains.
3 Enable structured interventions: Coherence metrics could guide improvements like chain-of-thought prompting, instruction tuning, or self-verification.

In the same way that coherence and correspondence may not capture the full spectrum of human judgment, evaluating the intrinsic properties of language models offers an important complement to these dimensions. For instance, Martin et al. (2021) propose assessing neural networks through heavy-tailed spectral properties of their weight matrices. These structural indicators have been shown to correlate strongly with generalization performance across models - even in the absence of traditional test data. By analyzing a model's internal structure, such methods offer a perspective that treats the model itself as data, complementing coherence-based evaluation with a view from the inside. This line of work reinforces our broader argument: that advancing LLM evaluation requires diverse and scalable approaches- those that assess both behavior externally and structure internally.

In short, just as the study of human cognition matured by expanding its understanding of rationality, LLM evaluation must move beyond narrow factuality checks. Accuracy does not speak with one voice—and understanding how models perform is central to grasping their capabilities and limitations. A dedicated, scalable coherence benchmark would mark an important step forward, as would further emerging criteria that focus on a model's internal characteristics.

3 Assessing LLMs through the lens of bounded rationality

Much like human cognition, LLMs operate under resource constraints. They must balance competing objectives—accuracy, latency, compute efficiency, and cost. This mirrors what Herbert Simon (1955) described as bounded rationality: the idea that decision-makers (including artificial systems) rarely have unlimited time or resources and therefore rely on heuristics to make "good enough" decisions under constraint, rather than always optimizing for perfect accuracy.

This framework offers a compelling analogy for how we should evaluate LLMs. While current evaluation metrics often emphasize static measures—such as factual correctness or performance on fixed tests—they typically ignore the computational trade-offs that define real-world deployment. For instance, high-performing models like GPT-4 Turbo or Anthropic's Claude 3 (Opus) may deliver excellent benchmark results, but they require vast GPU memory, distributed inference infrastructure, and expensive hardware acceleration. These systems are optimized for capability, not efficiency.

Meanwhile, smaller or more efficient models (e.g., Mistral-7B, DeepSeek-V2, or Phi-2) can deliver near-state-of-the-art performance on select tasks with significantly lower resource usage. In latency-sensitive applications (such as customer support or real-time decision aids), a slightly less accurate but immediate response may be more valuable than a more accurate yet delayed one.

The recent development of DeepSeek R2 in 2025 exemplifies this trade-off. Developed to be cost-effective and deployable on relatively constrained hardware, the model prioritizes throughput and latency over marginal gains in benchmark accuracy (Baptista et al., 2025). Similarly, new inference strategies like vLLM and GGUF-based quantization (e.g., running LLaMA-2 13B at 4-bit precision) show a growing interest in efficient deployment rather than leaderboard dominance.

Yet most public evaluation frameworks overlook these constraints, focusing almost exclusively on benchmark-based correctness. As a result, they fail to capture the resource–accuracy trade-off that is central to many applied AI systems. Just as bounded rationality urges us to assess human decision-making in light of ecological constraints, LLM evaluation should recognize that a model's real-world utility depends not only on what it gets right, but also on what it achieves within the limits of time, compute, and memory.

In short, the bounded rationality perspective invites us to ask different questions about LLMs: not only "How accurate is this model?" but also "How effective is it under pressure?" and "How well does it scale when resources are tight?" Incorporating such perspectives is crucial. Without it, LLM benchmarks risk promoting models that are academically impressive but operationally impractical.

4 Rethinking generality: lessons from ecological rationality

A longstanding critique from Gigerenzer and Todd (1999) is that many so-called cognitive "biases" identified by the heuristics-and-biases tradition arose from using abstract or ecologically invalid tasks. When tested in contexts that mirrored real-world decision-making—such as using natural frequencies instead of probabilities—many biases disappeared. This insight is highly relevant to today's conversations around LLMs: Are we evaluating these systems with benchmarks and tasks that reflect their intended real-world use.

The implications extend beyond benchmarking. The dominant narrative in AI assumes that generality is the hallmark of intelligence, with AGI (artificial general intelligence) as its ultimate form. But findings from ecological rationality and evolutionary psychology offer a different view: intelligence is about efficiency and adaptiveness within specific environments—not universal competence. Human cognition relies on specialized heuristics tailored to particular tasks and constraints. Similarly, recent trends in LLM research point toward a resurgence of task-specific optimization over raw generalization.

Concrete examples from the LLM landscape support this. For instance:

- Med-PaLM (Singhal et al., 2022) – a model fine-tuned on medical Q&A – outperforms general-purpose models like GPT-3.5 on domain-specific benchmarks such as USMLE-style exam questions.
- BloombergGPT (Wu et al., 2023), trained on a blend of financial news, filings, and proprietary data, significantly improves performance on finance-related NLP tasks compared to general models.
- WizardCoder (Xu et al., 2023) – a specialized coding assistant – can outperform a general LLM like ChatGPT on code generation and bug-fixing tasks.
- OpenAI's rumored *"Strawberry"* model (referred to unofficially by researchers) reportedly emphasizes logical consistency and chain-of-thought reasoning over general fluency, aiming to improve structured problem-solving.

Moreover, retrieval-augmented generation (RAG) architectures (Lewis et al., 2020) are increasingly used to bring domain-specific grounding into LLMs—especially in legal, medical, and enterprise contexts—underscoring the need for environment-aware adaptation.

These developments challenge the assumption that general-purpose models are universally superior. Instead, they highlight the importance of aligning model design, training, and evaluation with the ecological context in which models operate. Thus, just as ecological validity reshaped how we understand human reasoning, it should also reshape how we evaluate LLMs. Benchmarks must reflect context-specific demands, and success should be defined in terms of fit-for-purpose performance, not abstract generality. Without this shift, we risk misjudging the capabilities—and limitations—of these increasingly central AI systems.

5 Task redesign and structural interventions in LLM research

If the previous section raised concerns about representativeness and cross-task generalization, this one turns to robustness: Why do LLMs fail, and how can their outputs be systematically improved?

A central lesson from human judgment research is that performance can often be improved not by altering individual cognition directly, but by modifying the structure of the task or environment. This insight underpins the distinction between nudging and boosting—two families of interventions aimed at facilitating better decisions. Boosting, in particular, emphasizes durable, transparent improvements via structural changes to how information is presented (Hertwig and Grüne-Yanoff, 2017). A classic illustration comes from Gigerenzer and Hoffrage (1995), who showed that presenting statistical information as natural frequencies (e.g., "8 out of 10") rather than probabilities dramatically enhances diagnostic analysis. Such insights have informed practice in domains as varied as medicine, law, and public policy (Gigerenzer et al., 2007).

A similar structuralist perspective is emerging in LLM research. Interventions like prompt engineering, instruction tuning, and retrieval-augmented generation (RAG) have been shown to significantly improve model outputs without modifying the underlying weights. For example, Wei et al. (2022) demonstrated that well-designed prompts can elicit improved reasoning from models, at times rivaling the benefits of fine-tuning. RAG methods (Lewis et al., 2020) help mitigate hallucinations by grounding responses in external documents, while instruction tuning (Mishra et al., 2022) enhances alignment with task-specific requirements.

Crucially, these approaches are not merely engineering hacks—they benefit from being grounded in an understanding of the mechanisms underlying LLM errors. Zhang et al. (2024) offer a compelling case study, identifying knowledge overshadowing as a key driver of what they term amalgamated hallucinations. This phenomenon occurs when a model trained on exclusively true statements still produces incorrect outputs by conflating multiple factual patterns. The root cause is an imbalanced training distribution, where high-frequency conditions suppress lower-frequency—but equally valid—ones.

Their analysis yields three core insights:

- Systematic error patterns: Hallucinations follow predictable generalization dynamics, reflecting the statistical dominance of certain patterns in the training data.
- Causal structure: These error patterns emerge from biased token prediction conditioned by asymmetric exposure during training.
- Corrective strategies: A decoding technique known as *self-contrastive decoding* can offset these effects at inference time, without additional model retraining.

Zhang et al.'s work exemplifies what the philosopher Bechtel (2008) calls *mechanistic explanation*: identifying components, mapping their interactions, and designing interventions to influence outcomes. Rather than relying on anthropomorphic labels like "hallucination," their framework offers a clearer, system-level account of when and why certain failures emerge—and how they might be mitigated.

Still, these strategies have limits. As with boosting in human cognition, structural interventions do not eliminate foundational flaws; instead, they reshape inputs and task contexts to reduce error and enhance performance. That is precisely their strength. LLM task redesign, approached experimentally and informed by cognitive science, provides a principled way to test, probe, and refine model behavior. It enables us to study not just *what* models output, but how—and under what conditions—they succeed or fail.

6 LLM differences in thinking style

Another important lesson from research on human judgment comes from the study of individual differences. In particular, Stanovich's (2011) work on rational thought highlights the variability in how people reason—emphasizing distinctions in cognitive style, thinking dispositions, and the capacity for reflective override. Much of this builds on the heuristics-and-biases tradition, yet Stanovich's key contribution is to show that intelligence is not monolithic. He distinguishes between algorithmic-level intelligence (akin to IQ) and reflective-level rationality—the latter involving critical engagement with one's beliefs and goals.

This distinction offers a compelling analogy for understanding differences among LLMs. Just as people vary in their susceptibility to biases or their willingness to engage effortfully with complex problems, different LLMs exhibit distinct "thinking styles" shaped by their architectures, training regimes, and fine-tuning methods. Some may excel at structured reasoning (e.g., OpenAI's GPT-4), others shine in contextual interpretation (e.g., Claude 3.5), while still others trade raw capability for speed and deployability (e.g., Mistral L2 or DeepSeek-R1). Each model has its superpowers—and its blind spots.

This diversity matters for evaluation: a one-size-fits-all metric may fail to capture each model's unique strengths and weaknesses. Recent work by Martin et al. (2021) demonstrates that these behavioral differences are often reflected in a model's internal structure, revealing consistent patterns in how architectural and training choices shape model capabilities. Treating LLMs as interchangeable is as misleading as treating all human thinkers the same. Do different LLMs favor fluency over factuality? How do they respond under instruction pressure or in ambiguous contexts? Understanding and systematically comparing these tendencies - akin to studying cognitive styles in psychology - can help developers and users better match models to use cases and move toward a more principled science of evaluation.

7 Conclusion: toward an empirically grounded evaluation framework

Current LLM evaluation frameworks risk misalignment by over-relying on simplistic accuracy metrics and misleading metaphors. As argued throughout this paper, insights from human judgment research offer a pathway forward. Embracing lessons on heuristics, bounded rationality, and task design—while emphasizing mechanistic explanations, multi-dimensional accuracy models, and domain-sensitive evaluation strategies—can help build more robust evaluation frameworks for AI. By integrating such insights from cognitive science, AI assessment can evolve into a more rigorous, ecologically valid discipline, ensuring that future LLM development is driven by meaningful improvements rather than mere optimization for narrow benchmarks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AP: Conceptualization, Investigation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

Author AP was employed by company Coveo.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The author(s) verify and take full responsibility for the use of generative AI in the preparation of this manuscript. Generative AI was used in review Proofreading.

Publisher's note

References

Arkes, H. R., Gigerenzer, G., and Hertwig, R. (2016). How bad is incoherence? *Decision* 3, 20–39. doi: 10.1037/dec0000043

Baptista, E., Zhu, J., and Potkin, F. (2025). DeepSeek rushes to launch new AI model as China goes all in. Reuters. Available online at:https://www.reuters.com/technology/artificial-intelligence/deepseek-rushes-launch-new-ai-model-china-goes-all-2025-02-25/

Bechtel, W. (2008). Mental mechanisms: Philosophical perspectives on cognitive neuroscience. London: Routledge.

Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Front. Psychol.* 11:513474. doi: 10.3389/fpsyg.2020.513474

Brender, T. D. (2023). Medicine in the era of artificial intelligence: hey chatbot, write me an H&P. *JAMA Intern. Med.* 183, 507–508. doi: 10.1001/jamainternmed.2023.1832

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H., Kaplan, J., et al., (2021). Evaluating large language models trained on code. arXiv preprint. arXiv:2107.03374.

Chen, S., Zhao, Y., Zhang, J., Chern, I.-C., Gao, S., Liu, P., et al. (2023). FELM: Benchmarking factuality evaluation of large language models. *NeurIPS*. doi: 10.48550/arXiv.2310.00741

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Dean, J., et al., (2022). PaLM: scaling language modeling with pathways. *J. Mac. Learn. Res.* 24, 1–113.

Floridi, L., and Nobre, A. C. (2024). Anthropomorphising machines and Computerising minds: the Crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Mind. Mach.* 34. doi: 10.1007/s11023-024-09670-4

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Todd, P. M. (1999). Simple heuristics that make us smart. New York, NY: Oxford University Press.

Hammond, K. R. (2007). Beyond rationality: The search for wisdom in a troubled time: Oxford University Press.

Hertwig, R., and Grüne-Yanoff, T. (2017). Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986. doi: 10.1177/1745691617702496

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., et al., (2023). Swe-bench: Can language models resolve real-world github issues? arXiv preprint. arXiv:2310.06770.

Kahneman, D., and Tversky, A. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Katsikopoulos, K. V. (2009). Coherence and correspondence in engineering design: informing the conversation and connecting with judgment and decision-making research. *Judgm. Decis. Mak.* 4, 147–153. doi: 10.1017/S1930297500002588

Lewis, P., Perez, E., Piktus, A., Petrova, M., Goyal, N., Riedel, S., et alet al., (2020). Retrieval-augmented generation for knowledge-intensive NLP. *Adv. Neural Inf. Proces. Syst*. 33, 9459–9474. doi: 10.48550/arXiv.2005.11401

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 3214–3229).

Macmillan-Scott, O., and Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Soc. Open Sci*. 11:240255.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., et al. (2023). Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inform. Process. Syst*. 36, 46534–46594. doi: 10.48550/arXiv.2303.17651

Martin, C. H., and Mahoney, M. W. (2021). Implicit self-regularization in deep neural networks: evidence from random matrix theory. *J. Mach. Learn. Res.* 22:1. doi: 10.48550/arXiv.1810.01075

Martin, C. H., Peng, T., and Mahoney, M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nat. Commun.* 12:4421. doi: 10.1038/s41467-021-24025-8

Mishra, S., Khashabi, D., Baral, C., and McCallum, A. (2022). Cross-task generalization via natural language instructions. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 3493–3509).

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231

Polonioli, A. (2014). Blame it on the norm: the challenge from "adaptive rationality". *Philos. Soc. Sci.* 44, 131–150. doi: 10.1177/0048393113510468

Polonioli, A. (2015). The uses and abuses of the coherence–correspondence distinction. *Front. Psychol.* 6:1447. doi: 10.3389/fpsyg.2015.01447

Polonioli, A. (2016). Adaptive rationality, biases, and the heterogeneity hypothesis. *Rev. Philos. Psychol.* 7, 787–803. doi: 10.1007/s13164-015-0281-0

Silberling, A. (2024). Why AI can't spell "strawberry". TechCrunch. Available online at:https://techcrunch.com/2024/08/27/why-ai-cant-spell-strawberry/

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ*. 69, 99–118. doi: 10.2307/1884852

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., and Li, Z. (2022). Large language models encode clinical knowledge. *Natu*. 620, 172–180. doi: 10.1038/s41586-023-06291-2

Smith, A. L., Greaves, F., and Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Dig. Health* 2:e0000388. doi: 10.1371/journal.pdig.0000388

Stanovich, K. (2011). Rationality and the reflective mind. New York, NY: Oxford University Press.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Jegou, H., et al., (2023). LLaMA: open and efficient foundation language models. arXiv preprint. arXiv:2302.13971.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al., (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint. arXiv:2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Zhou, D., et al., (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advan. neural Inform. Process. Sys*. 35, 24824–24837.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., et al. (2023). BloombergGPT: a large language model for finance. arXiv preprint. arXiv:2303.17564.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., et al. (2023). WizardLM: Empowering large language models to follow complex instructions. arXiv:2304.12244v2.

Zhang, Y., Li, S., Liu, J., Yu, P., Fung, Y. R., Li, J., et al., (2024). Knowledge overshadowing causes amalgamated hallucination in large language models. Available online at: https://arxiv.org/html/2407.08039v1

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot arena. *Adv. Neural Inform. Process. Syst*.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. arXiv preprint. arXiv:2205.10625.