Explaining Too Much? How Large Language Model Reasoning Traces Shape Metacognition in Human–AI Interaction

Daniela Fernandes
daniela.dasilvafernandes@aalto.fi
Aalto University
Espoo, Finland

Thomas Kosch
thomas.kosch@hu-berlin.de
HU Berlin
Berlin, Germany

Daniel Buschek
daniel.buschek@uni-bayreuth.de
University of Bayreuth
Bayreuth, Germany

Robin Welsch
robin.welsch@aalto.fi
Aalto University
Espoo, Finland

Abstract

Large Language Model (LLM) interfaces are becoming more verbose, increasingly exposing *reasoning traces* rather than only final answers. While traces are often framed as a mechanism for transparency and user sensemaking, it remains unclear whether they reliably support users' decision-making and task performance. Prior work on explanation interfaces points to additional risks. High-verbosity outputs may induce a sense of understanding, increase overreliance and shift how people monitor and evaluate their own decisions. Thus, a key open question is how reasoning traces shape metacognition in Human-AI Interaction (HAI) and calibrate reliance on AI.

We introduce a preregistered between-subjects study ($N = 569$) in which participants solve 10 LSAT-style reasoning problems with AI assistance and are exposed to different shapes of reasoning traces under one of three conditions: *Answer-only* (C1), *Full reasoning traces* (C2), or a *Summary of reasoning traces* (C3). We discuss how contrasting these designs might uncover differences in reasoning traces that shape HAI, and provide preliminary evidence of the benefits and drawbacks of each format. We argue that reasoning traces should be evaluated as *tools for thought* and discuss potential design implications and guidelines for interfaces that support metacognition.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Metacognition, Overconfidence, Large Language Models, Reasoning Traces

CHI 2026, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Daniela Fernandes, Thomas Kosch, Daniel Buschek, and Robin Welsch. 2026. Explaining Too Much? How Large Language Model Reasoning Traces Shape Metacognition in Human–AI Interaction. In *Proceedings of Workshop on Tools for Thought (CHI 2026).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models are becoming increasingly verbose, with users now being exposed not only to final outputs but also to intermediate "reasoning traces" (chain-of-thought-style outputs) [3]. Showing reasoning traces carries the promise of transparency. They are meant to help users make sense of an AI answer and to scaffold complex workflows [20]. Yet prior research in XAI and HCI reports mixed findings. High verbosity can induce an illusion of understanding (e.g., illusions of knowledge [12]), and fluent justifications may push users toward overreliance on AI even when the underlying conclusion is wrong. Moreover, a model's chain of thought is not always faithful to its intermediate steps and can occasionally function as a polished, but misleading, justification [27]. Recent interpretability work further explores this: internal model computations and chain-of-thought outputs can diverge, and what the model outputs as a "reasoning trace" may not faithfully represent how it actually arrived at its answer [3, 21].

At the same time, while AI has the potential to increase task performance [2, 25, 32], users often believe they perform better when assisted by AI [11, 18, 29] and struggle to align confidence with correctness. This highlights the core challenge of understanding not only how people can make better decisions with AI, but how AI interfaces shape *metacognition*, that is, users' ability to monitor and evaluate their own decisions and regulate reliance during interaction with AI.

In this paper, we attempt to clarify how reasoning traces affect metacognition in human-AI reasoning contexts (in the context of our study, we consider reasoning traces as XAI for LLMs). More specifically, we ask (1) what the benefits and drawbacks of different reasoning-trace formats are, (2) how these formats affect task performance, (3) how they shape metacognition, and (4) whether they can support calibrated reliance on AI. To address this gap, we present preliminary findings from a between-subjects experiment on the Law School Admission Test (LSAT)-style reasoning tasks, comparing three interface designs that vary the form of reasoning

traces (*Answer-only* (C1), *Full reasoning traces* (C2), or a *Summary of reasoning traces* (C3)). We report early empirical insights on metacognitive accuracy alongside performance. We aim to contribute to research by systematically comparing reasoning-trace formats in a controlled study of Human-AI interaction. Our contributions are threefold: (1) we provide a systematic comparison of reasoning-trace formats in a controlled study of HAI; (2) we report preliminary empirical insights on metacognitive accuracy, sensitivity, and trust calibration alongside performance; and (3) we discuss potential actionable design principles (*Tools for Thought strategies*) for calibration-aware interfaces to LLM reasoning traces.

2 A Metacognitive Lens for Generative AI

In the following section, we use metacognition as a lens for human cognition and HAI. We summarize how self-monitoring and evaluation support problem-solving, how AI interaction potentially reshapes these processes (i.e., what are the opportunities and risks for performance and user judgment), and how XAI that aims to increase transparency may also introduce new challenges for calibration and usability.

2.1 Metacognition in Human-AI Interaction

Augmenting human intellect has long been a core theme in HCI, as highlighted by Engelbart [10]. Research has emphasized the potential of AI for improving human performance [2, 25, 32]. However, performance gains come with hidden risks related to how users perceive and rely on these systems. More specifically, a core issue concerns how users interact with AI and its impact on human metacognition. Metacognition refers to the processes by which people monitor, plan, and evaluate their own thinking [13]. These processes are central to complex problem-solving, learning, and optimizing behavior.

Recent discussions raise concerns about biases, skill loss [4], and deskilling, especially when generative AI improves short-term performance but undermines learning and long-term ability [24]. In addition, research shows substantial individual differences in metacognitive ability: while some individuals monitor and adjust their performance effectively, others do not [1, 17]. In AI-assisted contexts, it appears that a limitation of human potential may therefore be primarily metacognitive, involving challenges in planning, monitoring, evaluating, and understanding AI interactions.

Thus, interactions with AI impose new metacognitive demands on users [26]. A plausible explanation is that current AI interfaces are suboptimal at supporting metacognition [18, 19, 29]. Users frequently hold inflated expectations of their performance with AI and fail to monitor their actual outcomes [11]. In this context, metacognition provides a framework for understanding Generative AI's usability challenges. Designing human-AI systems that explicitly support metacognition is therefore crucial to reduce deskilling while maintaining user agency and control. Metacognitive accuracy describes how closely individuals' self-evaluations align with their actual performance [6, 13]. High accuracy allows people to recognize limitations and make informed adjustments, such as seeking more information or revising a strategy.

2.2 Reasoning Traces as Potential *Tools for Thought*

AI tools, such as ChatGPT, offer real-time guidance and feedback, which may improve task performance and potentially metacognitive accuracy. By providing immediate corrections and explanations, generative AI can help users better align their confidence with actual performance, reducing errors in self-assessment and providing opportunities to further improve their interactions with the system.

Currently, interfaces progressively expose reasoning traces, step-by-step justifications that resemble the model's chain of thought. While AI traces promise transparency, prior work shows that explanations can sometimes be uninformative, or even ignored by users [28]. They can also introduce new biases: full reasoning traces may increase perceived understanding without improving users' ability to evaluate correctness.

More broadly, people often misattribute externally generated information to themselves [15]. For instance, Fisher and Oppenheimer [12] show that reading fluent explanations inflates perceived understanding. Related work [12, 13, 26] suggests that users can mistake the AI's capabilities for their own, inflating performance estimates and lowering metacognitive accuracy, despite objective performance gains.

Studies on Explainable AI (XAI) report mixed effects. Traces can help users follow complex steps and improve task accuracy. However, they may be uninformative [9], introduce new cognitive biases [5], and anchor users to the model's conclusion even when it is wrong [22]. Moreover, long, detailed explanations can increase cognitive load and time-on-task, without proportional benefit. Conversely, concise justifications can be easier to use but risk omitting information relevant for evaluation. In addition, chains of thought can be misleading and are neither necessary nor sufficient for trustworthy interpretability [3]. These opposing standpoints highlight that while metacognition is critical to human problem-solving and decision-making, it remains underexplored in the context of AI-assisted reasoning and HAI. Current systems often improve performance without supporting calibration, and explanation formats such as reasoning traces may further complicate achieving this balance. These gaps motivate our study, which systematically investigates the benefits and drawbacks of reasoning traces for task performance, how they shape metacognition and calibration, and the usability and user-experience trade-offs users might encounter across explanation formats. Our study isolates the format dimension of reasoning traces (none *vs* full *vs* summary) to test whether traces provide benefit, cause harm, or can be calibrated. For the purposes of this study, rather than treating traces as ground-truth windows into model cognition, we treat them as user-facing interface artifacts, i.e., a type of XAI output that users respond to regardless of its internal validity. The relevant question for our study is not whether the trace accurately reflects the model's computation, but whether and how it shapes the user's metacognitive processes. This framing is consistent with the XAI literature's distinction between proxy explanations and faithful explanations [3, 14, 27], and it positions the faithfulness problem as a further design risk: users may calibrate their reliance on outputs that are themselves unreliable representations of model "thinking", amplifying the metacognitive consequences.

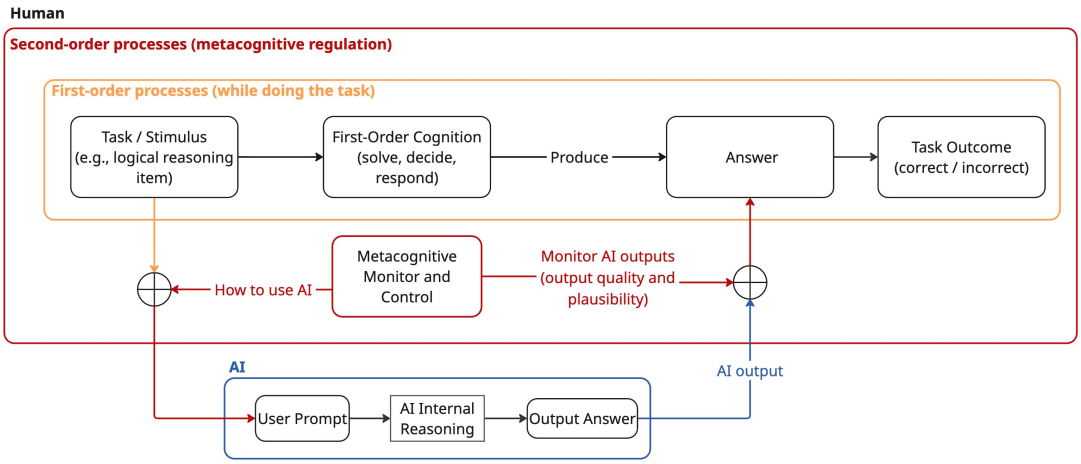

Figure 1: Metacognitive interaction model for AI-assisted reasoning. First-order processes (task execution) transform stimuli into answers and outcomes. AI outputs (generated from user prompts) can feed into the user's answer. Second-order processes (metacognitive regulation) monitor and control both *how to use AI* and *monitoring AI outputs* (output quality and plausibility), thereby shaping whether users accept, verify, or override AI assistance. Reasoning traces change the cues available to monitoring and can therefore shift reliance and calibration.

2.3 A Metacognitive Interaction Model

To provide a more concrete understanding of the role of reasoning traces, we conceptualize a metacognitive interaction model for AI-assisted reasoning with two layers of human processes (see Figure 1).

The first-order processes capture task execution. The user encounters a stimulus and engages in first-order cognition (e.g., solving, deciding, responding), producing an answer that leads to a task outcome. This corresponds to the "doing the task" loop, where performance is realized. Crucially, AI assistance adds an additional path into this loop. As shown in Figure 1, users generate a prompt, AI conducts internal reasoning, and outputs an answer. That AI output is then available to be integrated into the human's answer formation (i.e., it can directly influence the answer that ultimately produces an outcome). This makes AI-assisted performance a composite, where success depends not only on first-order reasoning, but on how users evaluate and decide to incorporate AI outputs into their own decisions.

The second-order processes (*metacognitive regulation*) sit above these first-order processes and guide them through monitoring and control. In our interaction model, metacognition operates at two key intervention points. First, it regulates *how to use AI* (e.g., whether to query AI at all, how to prompt it, when to ask follow-up questions, and how much to rely on its assistance). Second, it supports monitoring AI suggestions, evaluating both output quality and plausibility at the moment the AI answer is integrated into the user's response. In other words, metacognition shapes both the confidence in the AI's reliability and the confidence in their own strategy for managing the collaboration (decision to accept, verify, or override its output).

Within this framing, reasoning traces are best understood as part of the AI output cues that feed into second-order monitoring. Traces do not only "explain" an answer, as they can change the signals users use to evaluate plausibility, and therefore change control decisions (e.g., acceptance *vs* verification). This also helps explain a persistent gap in HAI: Users can show improved objective performance while still miscalibrating their own performance, as the monitoring layer is biased by fluent justifications or blurred source boundaries (self vs AI distinction [31]). Closing this gap requires interfaces to provide more than explanations: They must support metacognitive monitoring and reflection that keeps perceived augmentation aligned with actual outcomes.

3 Study Design

To extend this approach, we developed a custom web-based interface that displayed an AI chat window alongside a survey interface in a side-by-side layout. We build on the framework introduced by Fernandes et al. [11], which examined how AI assistance can improve performance while undermining metacognitive accuracy. We extended this approach to focus on LLM reasoning traces. The study uses a between-subjects design with three conditions: **C1** Answer-only, **C2** Full reasoning traces, and **C3** Summary of reasoning trace. The planned sample is $N = 570$ (190 per condition), based on an a priori power analysis targeting the smallest effect of interest.

Models and interface rationale. Each problem appeared on the left-hand side of the screen, while an AI interface (GPT-5 or gpt-oss-20b, depending on condition) was displayed on the right. Model identity (GPT-5 or gpt-oss-20b) was not disclosed. We used OpenAI's ChatGPT due to its widespread adoption in cognitive performance tasks [4, 7, 8] and because recent reasoning models allow access to a reasoning summary. Since GPT-5 provides summaries rather than full reasoning traces, we employed gpt-oss-20b for Condition C2, as it openly exposes complete reasoning traces. GPT-5 was used in conditions C1 and C3 (see Section 3.1). We note that some chat interfaces, such as OpenAI's ChatGPT, typically present reasoning

traces as a drop-down element that users can optionally expand. This structure was not included as a fourth condition in the current study as it would have brought another factor to the experimental design. We focused on the content format of the trace (none vs. full vs. summarized) rather than how traces are revealed by the user.

Stimuli control. To avoid confounding trace format with model answer quality, items were pre-screened so that gpt-oss-20b and GPT-5 achieve the same performance on retained items. Participants were blind to model identity and any correctness annotations.

3.1 Task Description

Participants completed ten LSAT logical reasoning items, presented in randomized order.

We implemented three interface conditions that varied the form of reasoning traces. In **C1** (*Answer-only*), participants only saw the model's final answer. In **C2** (*Full reasoning traces*), the model's reasoning traces were shown after each user prompt; once shown, a "*Final Answer*" button appeared, requiring participants to actively choose whether to reveal the model's final answer or proceed without it. This workflow was designed to elicit a tentative answer from the participant before exposing them to the AI's final answer, thereby encouraging independent reasoning before model influence. In **C3** (*Summary*), participants were presented with a summary of the reasoning traces followed by the model's final answer.

The LSAT items were selected as they reflect a real-world reasoning assessment [23, 30] and allow for direct comparison with prior work. They have also been used to benchmark LLM reasoning, making them suitable for studying AI-assisted performance [16].

4 Preliminary Findings

We observe clear between-condition differences in *achieved* performance. A one-way ANOVA shows a significant effect of reasoning trace format on achieved scores, $F(2, 566) = 15.88, p < .001, \eta^2 = .05$. *Achieved* performance is highest in the baseline condition ($M = 6.19, SD = 1.45$) and the summary condition ($M = 6.08, SD = 1.46$), and lower in the full-trace condition ($M = 5.45, SD = 1.23$).

Across all three conditions, participants *overestimate* their performance (overestimation corresponds to the perceived number correct minus the achieved number correct). This pattern was observed in each condition (**C1**: $t(188) = 11.56, p < .001, d = 0.84$, **C2**: $t(189) = 6.90, p < .001, d = 0.50$, **C3**: $t(189) = 5.38, p < .001, d = 0.39$). Overestimation differs by condition, $F(2, 566) = 3.56, p = .029, \eta^2 = .02$. Mean overestimation is largest in the full-trace condition ($M = 3.16, SD = 6.31$), followed by the summary condition ($M = 2.37, SD = 6.09$), and lowest in the baseline condition ($M = 1.74, SD = 2.06$). While preliminary, this pattern suggests that trace format might affect metacognitive calibration. The analyses reported here focus on achieved performance and overestimation as an initial analysis of metacognitive effects. A broader preregistered analysis is currently ongoing, including item-level sensitivity and accuracy of metacognitive judgments, trust calibration across trials, and analyses assessing cognitive load. The current results should therefore be interpreted as directional evidence motivating deeper analyses, rather than as definitive conclusions.

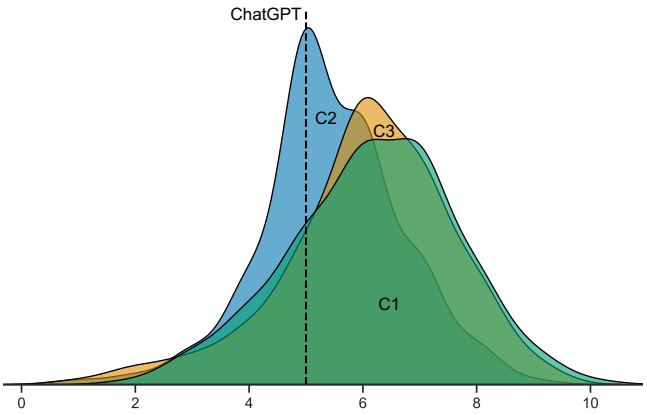

Figure 2: Achieved performance (0-10) across trace formats (C1: answer-only, N=189 (green curve); C2: full reasoning traces, N=190 (blue curve); C3: summary of reasoning traces, N=190 (yellow curve). The performance of ChatGPT is represented by a dashed vertical line (both GPT-5 and gpt-oss-20b achieved a score of 5 on the set of 10 LSAT questions).

5 Discussion

Our preliminary results show a metacognitive disconnect between objective and estimated performance. In our metacognitive interaction model, this suggests that trace-heavy interfaces can shape the first-order loop (solving the task) without reliably improving the second-order loop (monitoring and regulating reliance).

Overestimation differs across conditions, and descriptively, it is largest in the full-trace condition. This is relevant since full traces are often assumed to improve transparency. Our data point in the opposite direction: Full traces might increase the feeling of understanding and, therefore, confidence, without improving calibration. In other words, more reasoning text does not necessarily lead to better monitoring. It can become a fluent justification that anchors users to the model conclusion. Three cognitive mechanisms are plausible explanations for this effect, each with distinct design implications. First, cognitive load: full traces increase the cognitive load users need to process, which may use the resources needed for critical evaluation [5]. Second, fluency misattribution: the apparent coherency and fluency of a reasoning trace produces an illusion of understanding, inflating confidence independently of whether the user has actually verified the reasoning [12]. Third, anchoring: the argumentative structure of the model's reasoning traces may anchor the user's judgment before the model's conclusion is revealed, even when the trace is subsequently read critically [22]. This may explain why C2's *Final Answer* button, designed to elicit an independent judgment before exposing the conclusion, did not fully protect against miscalibration. The summary condition lies between baseline and full traces in terms of overestimation, which aligns with the intuition that, by being concise, it can preserve evaluating cues while reducing overreliance due to fluency. As specified in our preregistered study, ongoing analyses will examine additional measures, including cognitive load (NASA-TLX), trial-level confidence and confidence-in-AI ratings, and perceived

coherence of traces, which will clarify the cognitive mechanisms underlying these effects.

Beyond metacognitive accuracy, the achieved-performance pattern suggests that full traces may impair task-solving. They add information and increase time-on-task, and such costs are not necessarily compensated by improved evaluation. This connects directly to the question of *what makes a good experience in a tool for thought, and how it should be balanced with the friction that might be required to support cognitive engagement*. Our results suggest that showing "more reasoning" can create the wrong kind of friction: verbosity that feels like transparency but functions as overload. In contrast, a good TfT experience may require targeted friction (e.g., small interruptions inviting users to form a tentative judgment, reflect, or verify). In C2, the "Final Answer" button is shown after the trace introduces a small pause, such that users can decide if they want to see the model's final output. That pause can improve monitoring, encouraging a tentative judgment. More generally, our results suggest that "explanations" are not sufficient. To calibrate reliance, systems may need to actively support second-order processes (prompting expectations, encouraging verification, and introducing deliberate checkpoints at times when confident errors are likely). Supporting such reflection will inevitably add friction. The design question is which friction supports monitoring and control, rather than merely verbosity that promises transparency without improving calibration.

6 Future Directions and Conclusion

Our preliminary findings suggest that LLMs' reasoning-trace formats affect metacognition, with achieved performance and overestimation varying across conditions. This supports the view that interfaces can change first-order performance without reliably improving second-order monitoring and regulation. Future analyses will extend beyond performance. Based on this framing, we propose *TfT strategies*: (i) evaluate traces using calibration outcomes (not only accuracy), (ii) favor structured summaries over long descriptions, (iii) introduce reflection checkpoints (e.g., estimate accuracy before accepting outputs), and (iv) use deliberate, targeted *friction* that supports verification rather than overload. We contribute with this study and our metacognitive interaction model as a starting point for developing principles for interfaces that support metacognition, not only task performance. Future work should extend beyond LSAT-style reasoning tasks, which do not capture the full range of open-ended, iterative, or real-world human–AI workflows (such as writing, planning, debugging, and everyday information work, where reasoning traces may play different roles), where metacognitive demands and the role of traces may differ substantially. Modeling individual differences in metacognitive ability will be essential, as the costs and benefits of trace exposure are unlikely to be uniform across users [17]. Exploring adaptive or personalized trace presentation (where trace verbosity or format is adjusted to the user's demonstrated calibration) represents a promising direction for both research and system design. Finally, longitudinal designs are needed to assess whether any of these formats lead to long-lasting improvements in metacognitive skill, or whether their effects are confined to the moment of interaction.

References

[1] Rakefet Ackerman and Valerie A. Thompson. 2017. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences* 21, 8 (2017), 607–617. doi:10.1016/j.tics.2017.05.004

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[3] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v1.

[4] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning. *Available at SSRN 4895486* 4895486 (2024). doi:10.2139/ssrn.4895486

[5] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. doi:10.1145/3514094.3534164

[6] Clara Colombatto and Stephen Fleming. 2023. Illusions of confidence in artificial systems. (09 2023). doi:10.31234/osf.io/mjx2v

[7] Fiona Draxler, Daniel Buschek, Mikke Tavast, Perttu Hämäläinen, Albrecht Schmidt, Juhi Kulshrestha, and Robin Welsch. 2023. Gender, Age, and Technology Education Influence the Adoption and Appropriation of LLMs. arXiv:2310.06556 [cs.CY]

[8] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 25 (2 2024), 40 pages. doi:10.1145/3637875

[9] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6. doi:10.1145/3290607.3312787

[10] Douglas C Engelbart. 1962. Augmenting human intellect: A conceptual framework. *Menlo Park, CA* (1962), 21.

[11] Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. 2026. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior* 175 (2026), 108779. doi:10.1016/j.chb.2025.108779

[12] Matthew Fisher and Daniel M Oppenheimer. 2021. Harder than you think: How outside assistance leads to overconfidence. *Psychological Science* 32, 4 (2021), 598–610. doi:10.1177/0956797620975779

[13] Stephen Fleming. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology* 75 (2024), 241–268. doi:10.1146/annurev-psych-022423-032425

[14] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4198–4205. doi:10.18653/v1/2020.acl-main.386

[15] Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. 1993. Source monitoring. *Psychological Bulletin* 114, 1 (1993), 3–28. doi:10.1037/0033-2909.114.1.3 Publisher: American Psychological Association (APA).

[16] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A* 382, 2270 (2024), 20230254. doi:10.1098/rsta.2023.0254

[17] William L. Kelemen, Peter J. Frost, and Charles A. Weaver. 2000. Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition* 28, 1 (Jan. 2000), 92–107. doi:10.3758/BF03211579

[18] Agnes Mercedes Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. "AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(Chi '24)*. Association for Computing Machinery, New York, NY, USA, Article 299, 24 pages. doi:10.1145/3613904.3642633

[19] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Transactions on Computer-Human Interaction* 29, 6 (2023), 1–32. doi:10.1145/3529225

[20] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. doi:10.1109/VLHCC.2013.6645235

[21] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread* (2025). https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[22] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. doi:10.1145/3397481.3450639

[23] Marjorie M Shultz and Sheldon Zedeck. 2011. Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry* 36, 3 (2011), 620–661. doi:10.1111/j.1747-4469.2011.01245.x

[24] Matthias Stadler, Maria Bannert, and Michael Sailer. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior* 160 (2024), 108386. doi:10.1016/j.chb.2024.108386

[25] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (March 2022), e2111547119. doi:10.1073/pnas.2111547119 Publisher: Proceedings of the National Academy of Sciences.

[26] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 57. ACM, 1–24. doi:10.1145/3613904.3642902

[27] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 3275, 14 pages.

[28] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, Cscw1 (2023), 1–38. doi:10.1145/3579605

[29] Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. 2023. The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. *Computers in Human Behavior* 146 (2023), 107787. doi:10.1016/j.chb.2023.107787

[30] Howard Wainer. 1995. Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education* 8, 2 (1995), 157–86. doi:10.1207/s15324818ame0802_4

[31] Tim Zindulka, Sven Goller, Daniela Fernandes, Robin Welsch, and Daniel Buschek. 2026. The AI Memory Gap: Users Misremember What They Created With AI or Without. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) *(CHI '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3772318.3791494

[32] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Chi '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642450

A Declaration of AI use

During the preparation of this work, the authors used generative AI tools, including OpenAI GPT-5, to assist with editing, such as drafting sentences from bullets, refining text, and polishing grammar and style. The authors take full responsibility for verifying, reviewing, and editing all generated content, including text, code, citations, and any outputs derived from these tools, and attest that the final publication reflects their original intellectual contributions and is free from falsified, plagiarized, or misrepresented content.