

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

Main theme(s): dTTF Outcomes: Definition and Measurement & TTF Experience and Adoption

Target domain(s): Reasoning and decision-making

Cognitive 'target(s)': Metacognitive monitoring and regulation; reasoning, critical thinking

Type of contribution & main idea

The effect of reasoning traces on decision-making is unclear (they might not contribute to "better transparency"). In our study, traces format affect **achieved performance and overestimation**. This suggests traces can shape performance without improving self-monitoring, interface should be designed and evaluated as TTF that aim to improve calibration, not only performance.

The AI tool's key characteristics

We implemented three interface conditions that varied the form of reasoning traces. Each problem appeared on the left hand side of the screen, while an AI interface (GPT-5 or gpt-5o), depending on condition) was displayed on the right.

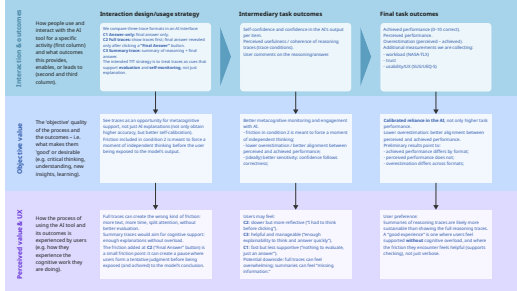
In C1 (Answer-only), participants were only presented with the model's final answer.

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.

What you would like to discuss

When do reasoning traces might harm metacognition? What would a "good" trace format that supports metacognition? How should this balance good experience with the friction required for cognitive engagement? What measures best capture "thinking better" beyond accuracy?



What would you like to take away from the workshop?

Discuss on a metacognitive interaction model as a starting point for developing principles for interfaces that support metacognition, not only task performance.

Key references (e.g. of main theories, empirical evidence, measurement methods etc.)

[1] Daniela Fernandes, Steven Villa, Saba Noor, Olya Haviv, Carol Bucher, Alberto Schmidt, Thomas Koch, Chenshan Shen, and Robin Strub. 2023. AI makes you smarter but more the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior* 175 (2023), 108775. doi:10.1016/j.chb.2023.108775 [2] Lev Tsetserlin, Viktor Klemov, Avner Shmida, Ana Eladshin-Shefi, Adam Serfaty, Nargal Serfaty, and Leon Ronen. 2024. The Metacognitive Benefits and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 1, ACM, 5–24. doi:10.1145/3695.842262 [3] Matthew Fisher and Daniel M. Oppenheimer. 2021. Healer: How you think you think. *How outside assistance leads to overconfidence*. *Psychological Science* 32, 6 (2021), 108–110. doi:10.1177/0956797620975779 [4] Paul Biane, Ting-Yu Wu, Ishita Dasgupta, Michael Lee, Wee-Hong Ng, Noah Singer, Nicolas Collignon, Chenshan Shen, Isabelle Lee, Atsushi Kawan, et al. 2023. Chat-ai thought is not explainability. *Preprint, arXiv:2312.14711*, v1.

How do people's goals 'interact with' the AI tool's goals? Is there a tension?

Short-term goal (user goal): get the right answer fast, feel confident and reduce effort.

Long-term goal (user goal): understand their own reasoning limits; learn and build better judgment about when to trust AI.

AI tool goal: produce coherent, fluent, convincing responses (not, with traces, coherent justifications).

Tension: traces can make the system feel more transparent and trustworthy, but that same fluency can **obscure subjective understanding** and also reliance on the AI, even when it doesn't improve task accuracy. This creates a calibration gap (and the "worry friction" problem).

How do you expect people to continue using this AI tool?

Users will likely adopt the trace formats that feel helpful with minimal effort. The summary format is more sustainable than full traces, as it requires less effort. Continued use depends on whether traces consistently provide value without increasing workload. Ideally, the traces would match the user needs and task difficulty.

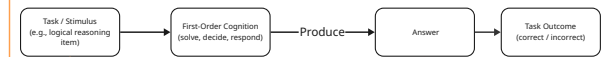
How to proceed with this work/idea?

Next, we will complete the preregistered analyses and iterate on trace-based interface designs that explicitly support **second-order processes** rather than adding verbosity. (i) evaluate traces using calibration outcomes (not only accuracy); (ii) build structured summaries; user long descriptions; (iii) introduce reflection (checkpoints) to e.g., estimate accuracy before accepting outputs; (iv) use deliberate, targeted friction that supports verification (rather than overload).

Human

Second-order processes (metacognitive regulation)

First-order processes (while doing the task)



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.