Qeios

Peer Review

Review of: "Is DeepSeek a Metacognition AI?"

Marek Urban[1,2]

1. CREAITE Lab, Institute of Psychology, Brno, Czechia; 2. Institute for Research in Social Communication, Bratislava, Slovakia

The central question of the article is intriguing: can a reasoning model acquire metacognitive abilities purely through reinforcement learning (RL)? The author builds on claims made in the DeepSeek-AI (2025) model card, where the developers state that the model was not explicitly trained on problem-solving strategies, yet it developed such abilities during RL. However, several issues arise with these claims.

First, while in humans the use of “metacognitive” language often signals underlying metacognitive processes, it is unclear whether similar language use in large language models (LLMs) reflects genuine metacognitive operations. That is, we do not know whether the model has developed a metacognitive “layer” over its cognitive processes or is merely generating words that are stylistically expected (or statistically appropriate) in such contexts. LLMs may simply mimic the way humans talk about problem-solving without engaging in the underlying processes themselves. This concern aligns with findings from Anthropic’s paper *Reasoning Models Don’t Always Say What They Think* (Chen et al., 2025), which shows that chain-of-thought reasoning often fails to reflect the actual computations behind the model's answers.

Second, in human psychology, metacognition involves both calibrated monitoring—knowing when one is likely right or wrong—and strategic control—choosing how to act based on that awareness. The article, however, presents no evidence for either of these mechanisms. It lacks confidence-calibration metrics (e.g., Bias Index), demonstrations of selectively timed halt-or-continue decisions (such as spending more time on uncertain tasks), and indications of cross-task transfer of self-monitoring abilities. Consequently, the behaviors described may represent superficial mimicry rather than functional metacognition. I recommend the authors engage with findings from Griot et al. (2025), *Large Language Models Lack*

 doi.org/10.32388/4GLV4G

Essential Metacognition for Reliable Medical Reasoning. While that study did not include reasoning models per se, its benchmarks could be fruitfully extended to evaluate them.

That said, as noted at the outset, the idea that reasoning models might become metacognitive is compelling for several reasons. First, it suggests that metacognition may be an emergent property of advanced learning systems. Second, it raises questions about how such models will evolve when AI developers begin explicitly training them on expert problem-solving strategies, as envisioned by Kokotajlo et al. (2025) in *AI 2027*. Third—and perhaps most consequential—it implies that future language model development may increasingly require collaboration with cognitive and educational psychologists to guide the training of models toward more robust and reliable forms of reasoning.

References:

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning models don't always say what they think. Anthropic. https://www.anthropic.com/research/reasoning-models-dont-say-think

DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. https://doi.org/10.48550/arXiv.2501.12948

Griot, J., Mahmood, T., Li, Y., Singh, R., & Doshi-Velez, F. (2025). Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*. 16, article number 642. https://doi.org/10.1038/s41467-024-55628-6

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). *AI 2027*. https://www.ai-2027.com/

Declarations

Potential competing interests: No potential competing interests to declare.