

Taking a Language Model’s Word for It?

How people check answers from an AI chatbot and from an encyclopedia

Xiaochuan Li University of Technology Nuremberg · M.Sc. Human and AI



Full report

Take-home. People who could check with an AI chatbot first did not double-check less; descriptively, they double-checked more. Neither group difference was reliable ($N = 47$).

Introduction

Language models now answer many everyday questions. Whether people let such an answer settle a question shapes how information passes between AI systems and humans.

ChatGPT is discussed as an *epistemic authority*: a source whose answer counts as a reason to believe (Cooper et al., 2024; Hauswald, 2025). Relying on it is a form of cognitive offloading (Shanmugasundaram & Tamilarasu, 2023).

Does checking with an AI chatbot first, rather than an encyclopedia, change how often people check their answers, and how often they check again elsewhere?

- H1** With ChatGPT first, people check **more** answers in round 1.
- H2** With ChatGPT first, people check **fewer** answers with the second source.

Method

- Design** Between subjects; first tool assigned at random: **Wikipedia first** or **ChatGPT first**.
- Sample** 49 took part, 2 declined consent; 47 analysed (29 and 18). Convenience sample.
- Task** Ten two-option general-knowledge questions (*What is the coloured portion of the eye called? Iris / Pupil*). Each tool’s answer appeared as a fixed screenshot.

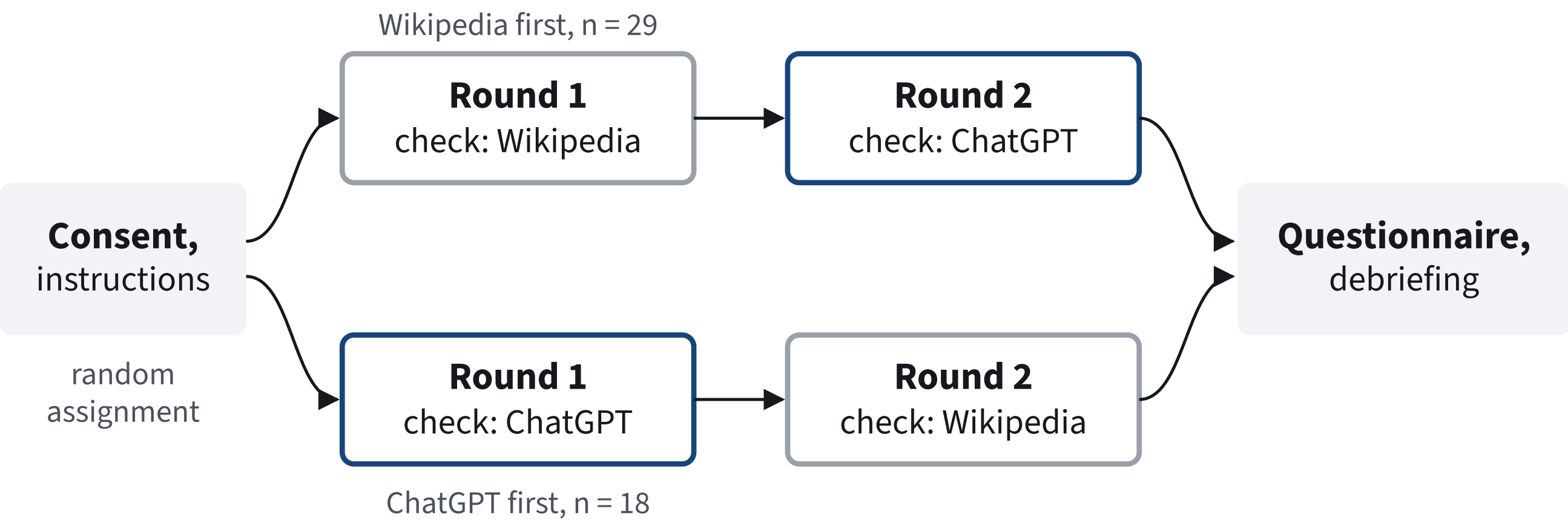


Figure 1. Every answer could be checked twice: first with the assigned tool, then with the other. Participants also rated their certainty (1–4).

- Measures** Answers checked in round 1 and in round 2 (each 0–10); certainty; accuracy.
- Analysis** Welch’s t -tests, Cohen’s d , Pearson correlations. Unanswered check prompts count as not checked.

Results: source order and checking

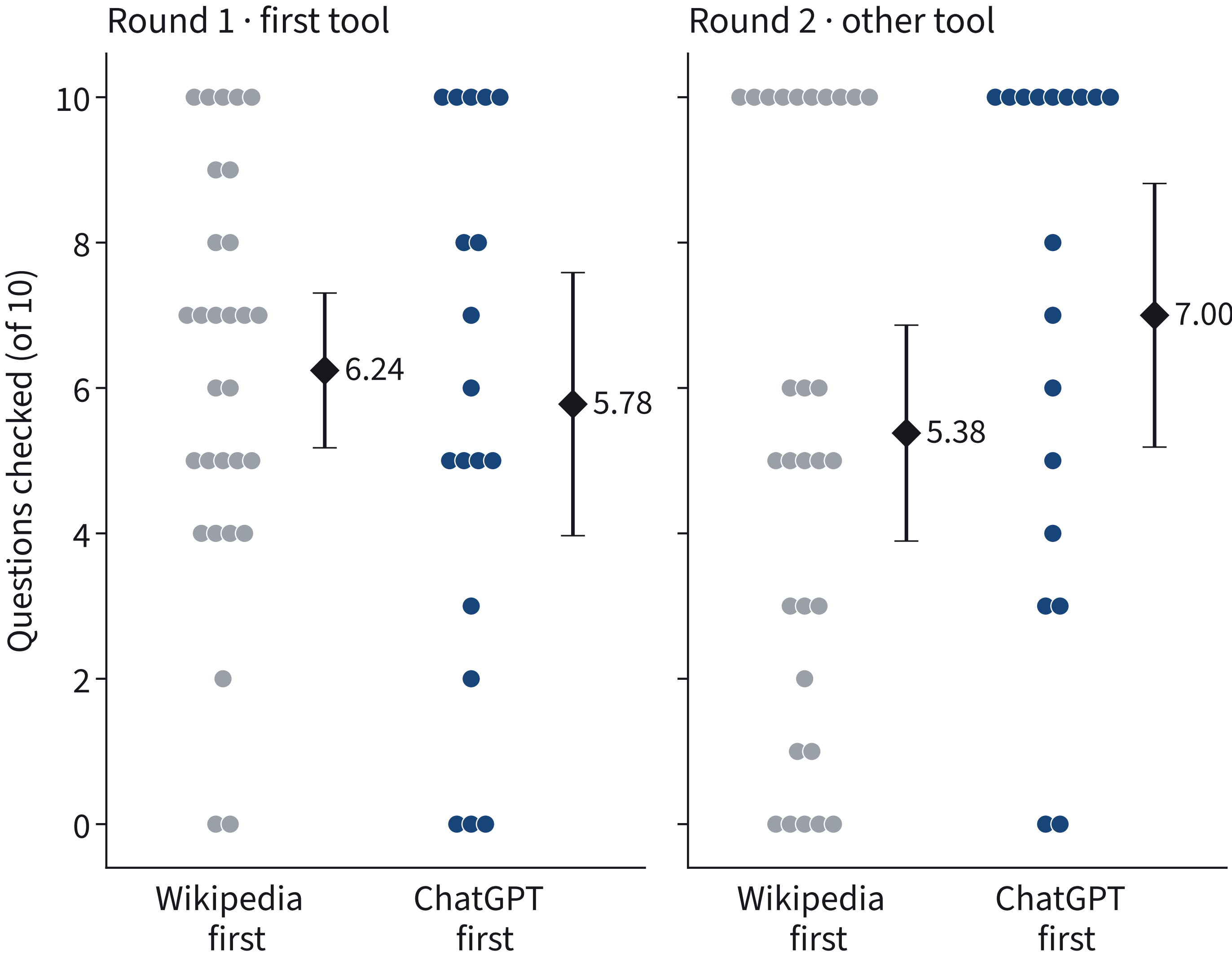


Figure 2. ChatGPT-first participants checked slightly less in round 1 and more in round 2. Dots: participants; diamonds: mean with 95% CI.

	Wikipedia first ($n = 29$)	ChatGPT first ($n = 18$)	Welch t (df)	p	d
Round 1	6.24 (2.80)	5.78 (3.64)	−0.46 (29.4)	.65	−0.15
Round 2	5.38 (3.90)	7.00 (3.65)	1.44 (38.1)	.16	0.43

Table 1. Answers checked, M (SD) of 10. t and d : ChatGPT first minus Wikipedia first.

- H1 Not supported.** ChatGPT first led to slightly *fewer* round-1 checks.
- H2 Not supported.** ChatGPT first led to *more* round-2 checks; 9 of 18 checked nine or more answers.

Round-2 counts split between almost none and almost all.

Does doubt drive double-checking?

If it does, round-2 checks should rise as accuracy and certainty fall.

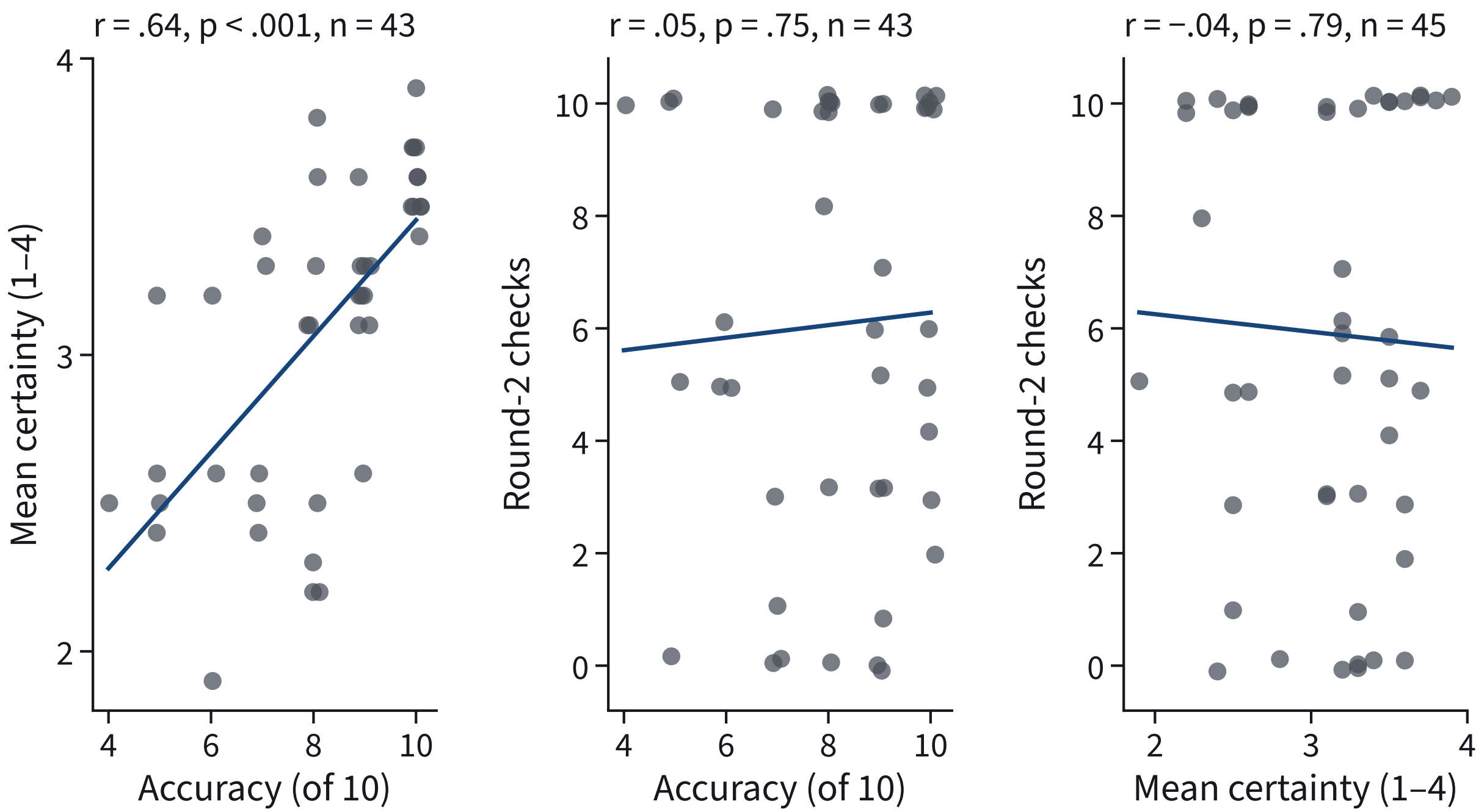


Figure 3. Certainty tracked accuracy, but neither predicted round-2 checking. Pairwise complete cases; points jittered.

Conclusions

- An AI-first check did not reduce double-checking; descriptively, more followed.
- If this holds in larger samples, AI answers prompt further checking rather than settle the question.
- Double-checking did not follow doubt; one-click checks may explain how often people checked.

Limitations and next steps

Small, unbalanced sample; easy items; cost-free checks. Next: a larger sample, harder items, checks that cost time, and item-level models of certainty and trust in each tool.

Cooper, G., Tang, K.-S., & Rappa, N. A. (2024). Generative artificial intelligence as epistemic authority? In H. Crompton & D. Burke (Eds.), *Artificial Intelligence Applications in Higher Education*. Routledge.
Hauswald, R. (2025). Artificial epistemic authorities. *Social Epistemology*, 39(6), 716–725.
Shanmugasundaram, M., & Tamilarasu, A. (2023). The impact of digital technology, social media, and artificial intelligence on cognitive functions: A review. *Frontiers in Cognition*, 2, 1203077.