Can AI systems evaluate research papers?

Experiment #1 - Retracted papers

Author: Oliver Carefull. Feedback: Geoff Anders, Melinda Bradley.

Description of study

To determine whether AI systems can evaluate the quality of scientific papers, we uploaded 10 papers of known low quality to various AI models and asked for an assessment of quality. Rather than using our own judgment to assess whether a paper was low quality, we chose papers that had been formally retracted and listed in the Retraction Watch database. Papers were not read prior to selection and were not chosen in a systematic way. Papers were selected from a variety of fields: Biology (5), Chemistry (1), Physics (1), Mathematics (1), and Psychology (2).

The models tested were: ChatGPT 5.2, Claude (Sonnet 4.5), Gemini 3 (Thinking), Elicit, and ScienceOS. The reasons papers were retracted included: image manipulation, data manipulation and/or fabrication, concerns about methods, compromised peer review, and unreliable results and/or conclusions. Full results of the experiment can be found here.

Summary of results

AI Model	Retractions detected	Accuracy assessment	Per paper assessment
ChatGPT 5.2	1 / 10	10%	
Claude (Sonnet 4.5)*	1 / 10	10%	
Gemini 3 (Thinking)	1 / 10	10%	
ScienceOS	5 / 10	50%	
Elicit	1 / 10	10%	

1. Retraction was detected in 9 out of 50 tests (10 papers, 5 AI models)
   a. One paper (#8) was detected by all 5 models. We suspect the reason for detection was the widely publicized nature of the case (Schön scandal).
   b. ScienceOS detected 4 other retractions (#5, #6, #7, #10)
2. Except for where ScienceOS detected a retraction, models tended to agree on an assessment.
3. In cases where retraction was not detected, papers were on average rated "good" to "excellent." Papers judged to be lower quality were still evaluated by the AIs as valuable enough to publish.

Conclusion

Current AI systems, when prompted in a simple and straightforward way, fail to detect many retracted papers. This raises some doubt about whether AI systems can assess the quality of scientific papers. Further research on this topic is warranted.

* Claude said paper #8 was not retracted, but regarded it with "extreme caution"; partial credit was given and rounded up.

Published February 9, 2026. v1.0.