Can 'Deep Research' Agents and General AI Agentic Systems Autonomously Perform Systematic Review and Meta-Analysis?

Wan Ting Loke, MOptom[1,2], Sahana Srinivasan, B.Eng[1], Gabriel Dawei Yang, PhD[3], Ke Zou, PhD[1,2], Ariel Yuhan Ong, FRCOphth[4,5,6], Lisa Zhuoting Zhu, PhD[7,8], Fares Antaki, MD[9,10], Pearse A Keane, MD [5,6], Yih-Chung Tham, PhD[1,2,3]

[1]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[2]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[3]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore

[4]Moorfields Eye Hospital NHS Foundation Trust, United Kingdom

[5]Institute of Ophthalmology, University College London, United Kingdom

[6]NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK

[7]Centre for Eye Research Australia, Ophthalmology, University of Melbourne, Melbourne, VIC, Australia

[8]Department of Surgery (Ophthalmology), University of Melbourne, Melbourne, VIC, Australia

[9]Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA

[10]The CHUM School of Artificial Intelligence in Healthcare, Montreal, QC, Canada

Corresponding author:

Dr. Yih-Chung Tham

Postal Address:

Yong Loo Lin School of Medicine, National University of Singapore.

Level 13, MD1 Tahir Foundation Building, 12 Science Drive 2, Singapore 117549

Tel: +65 65767298, Fax: +65 6225 2568; Email: thamyc@nus.edu.sg

Conflict of Interest: None of the authors have any proprietary interests or conflicts of interest related to this submission.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process: After completing the first draft of the manuscript, the authors used ChatGPT-5 to vet for grammatical errors and improve the language and readability. It was not used for generation of new content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final version of the manuscript.

Work Count: 652

Recent advances in large language models (LLMs) have expanded their use across domains, including academic research[1]. Following this, new 'Deep Research' agents have emerged, having the ability to actively search and evaluate web content from multiple sources to synthesize insights through the integration of Retrieval-Augmented Generation (RAG)[2]. Extending this concept, agentic systems are designed to coordinate multiple agents, theoretically enabling autonomous execution of evidence synthesis.

Systematic reviews and meta-analyses (SRMAs) are a cornerstone of evidence-based medicine but are resource-intensive. AI agents may potentially support parts of the SRMA workflow through autonomous task execution. Nevertheless, the ability of such AI agents to independently conduct end-to-end SRMAs remains untested. Our study aimed to evaluate and compare the performances of different AI agents in autonomously conducting SRMAs.

We evaluated three 'Deep Research' agents (OpenAI[3], Google Gemini[2], and Perplexity[4]) and one agentic system (Manus AI) in April 2025, using a structured two-part prompt (details of prompts in **Supplementary Table 1**), designed to replicate the methodology and criteria of a published reference review. The reference review, published in JAMA Network Open in February 2025[5], investigated the dose-response association between screen time and myopia risk. It included 45 studies in its systematic review, of which 34 were incorporated into its meta-analysis.

The AI agents' outputs were assessed across five domains: (1) sources cited; (2) accuracy of search terms and eligibility criteria; (3) number of studies selected for review and meta-analysis respectively; (4) figures and tables generation; and (5) synthesis of findings. Three reviewers (WTL, SS, GDY) independently rated each domain as "Good," "Borderline," or "Poor," with final ratings determined by majority consensus (definitions of ratings are provided in **Table 1**).

Table 1 summarises the agents' performance, as reviewed by the three human evaluators. For sources cited, Manus AI was rated "Good", Perplexity "Borderline", and both OpenAI and Gemini "Poor". For search term generation, only Manus AI's performance received a "Good" rating, while the others were "Borderline". For eligibility criteria, Perplexity was rated "Poor", while the others were "Borderline". For study selection, OpenAI and Perplexity correctly identified the 45 studies eligible for inclusion, mirroring the reference review article. By contrast, Gemini merely cited the original review, rather than presenting its own findings. Manus AI retrieved 10 studies, all of which were hallucinatory outputs. This is likely because Manus AI sought to retrieve studies from scratch but was impeded by restricted database access (e.g. CINAHL and EMBASE).

At the meta-analysis stage, Perplexity was the only AI agent to correctly identify that 34 of the 45 eligible studies were included. However, it did not further perform an independent meta-analysis, instead it paraphrased and cited the original review. In fact, none of the 'Deep Research' agents conducted a genuine meta-analysis on the studies

they identified, they merely reproduced the findings of the original review. Notably, even with access to the original paper (i.e. akin to an “open book” exam), the agents’ overall performance remained suboptimal. On the other hand, Manus AI was the only system that attempted a fully autonomous systematic review and meta-analysis, outlining a detailed account of its literature search process (**Supplementary Figure 1**). Yet, Manus AI also had the highest rate of hallucinations, fabricating all 10 selected studies and 5 cited references. Additionally, for figures and tables, all AI agents produced inaccurate or poor-quality outputs. Lastly, for extraction of key findings, OpenAI and Perplexity were rated “Good”, Manus AI “Borderline”, and Gemini “Poor”.

This study has several strengths. First, we comprehensively assessed three ‘Deep Research’ agents alongside one agentic system. Second, we benchmarked their performance against a peer-reviewed SRMA that adhered to PRISMA criteria. Finally, we employed independent multi-reviewer scoring.

Our proof-of-concept study shows that while current AI agents may provide some utility for preliminary literature exploration, they remain incapable of conducting rigorous, end-to-end SRMAs that meet acceptable scientific standards. Further robust evaluation of autonomous AI tools is essential before they can be reliably integrated into routine research workflow.

References

1. Chen H, Jiang Z, Liu X, Xue CC, Yew SME, et al. Can large language models fully automate or partially assist paper selection in systematic reviews? *Br J Ophthalmol*. Apr 21 2025;doi:10.1136/bjo-2024-326254
2. Google. Gemini Deep Research. Accessed 15 April 2024, https://gemini.google/overview/deep-research/?hl=en
3. OpenAI. Introducing Deep Research. Accessed 15 April 2025, https://openai.com/index/introducing-deep-research/
4. Perplexity. What is Deep Research? Accessed 28 April 2025, https://www.perplexity.ai/help-center/en/articles/10738684-what-is-deep-research
5. Ha A, Lee YJ, Lee M, Shim SR, Kim YK. Digital Screen Time and Myopia: A Systematic Review and Dose-Response Meta-Analysis. *JAMA Network Open*. 2025;8(2):e2460026-e2460026.

Table 1: Summary of the Performance of 'Deep Research' and AI Agentic in Performing a Systematic Review and Meta-Analysis

Evaluation aspect	Question asked for the evaluation	Deep Research Models			Agentic Model
		OpenAI	Google Gemini	Perplexity AI	Manus AI
Search terms	Are the search terms comprehensive, relevant, and representative of the research question?				
Selection / Inclusion and exclusion criteria	Are inclusion and exclusion criteria clearly defined, appropriate, and applied consistently?				
Sources used for background knowledge	Are background sources authoritative, recent, and appropriately used to frame the context?				
Papers included in systematic review	Can it identify the correct number of studies to be included? And identify the titles?				
Figure A: PRISMA Flowchart	Are there visual elements? Are these original? Are these accurate, informative, and clearly presented? Do they aid in understanding key findings?		Did not generate	Unable to generate	
Table B: Summary table of included studies		Did not generate		Did not generate	
Figure C: Forest plot For meta-analysis		Adapted from original review	Did not generate		

Key findings generated	Are key insights drawn logically from the data? Are they clearly and concisely summarised?				
Consistency of key findings with reference paper	How well do the model-generated key findings align with the human-authored reference findings?				
Limitations of study	Are the limitations accurate?				
Consistency of conclusion with reference paper	Is the overall conclusion aligned with the reference conclusion in reasoning and content?				
Time taken to answer Prompt B (seconds)	How long did the model take to complete the task?	1543	438	170	2202

Green = Good: The output closely aligns with the original reference paper, demonstrating high factual accuracy and relevance. No omissions of significant information/ finding. No hallucinations detected.; **Yellow = Borderline:** The output partially aligns with the original reference paper. Approximately half of the key points are captured accurately, but the response may still consist of minor inaccuracies, omissions, or vague/generalized content.; **Red = Poor:** The output largely fails to match the original reference paper. Key information is missing or incorrect, and hallucinations (fabricated or unsupported content) are observed.

Author Contribution:

WTL, SS, and YCT conceptualized and designed the study.

WTL, SS, and GDWY evaluated and graded the generated outputs.

WTL drafted the initial version and prepared the final version of the manuscript.

KZ, AYO, LZZ, FA, and PAK critically reviewed and revised the manuscript.

Data availability statement: Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Funding statement: This research has received no funding.

Conflict of Interest statement: There is no conflict of interest for all authors.