Indirect Source Referencing by Medical Large Language Models

Bar Borohov[1,2], Alexandra Rabotin[3]

[1]Rutgers Biomedical and Health Sciences, Rutgers University, New Jersey; [2]School of Public Health, Rutgers University, New Jersey;
[3]Mission Community Hospital, Panorama City, California

RUTGERS HEALTH
Environmental and Occupational Health Sciences Institute

Introduction

- Earlier GPT models (3.5 & 4) were shown to produce fabricated literature and citation errors in medical literature reviews.[1]
- **Indirect source referencing** is a type of inaccurate citation where authors cite a secondary source instead of the primary study.[2]
- Prior studies reported a 6-21% rate of indirect source referencing.[3,4]
- We aimed to evaluate how often two commercially available medical Large Language Models (LLMs) use indirect source referencing.

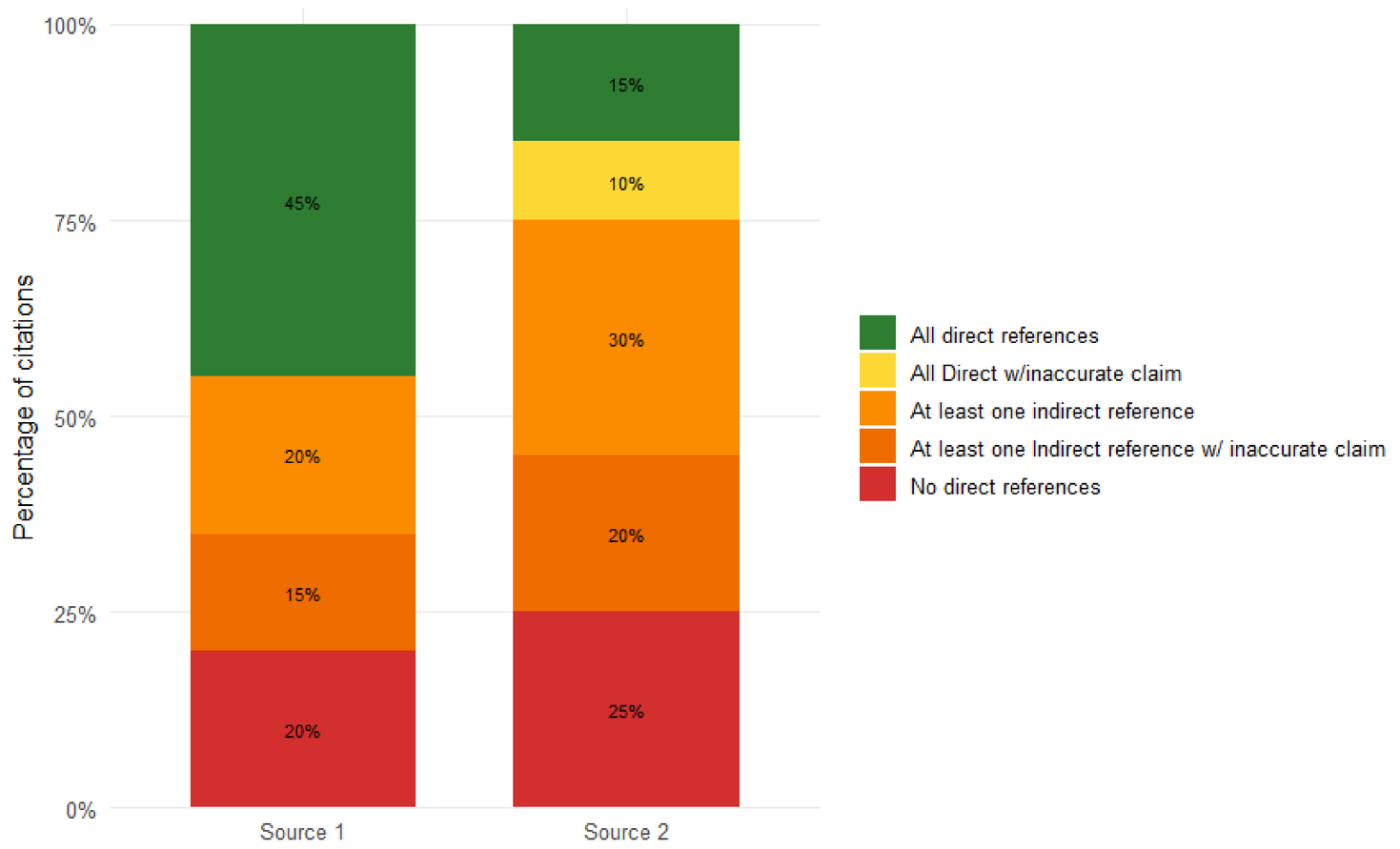

Figure 1. Directness and Accuracy of References by LLM Source (n=20 each)

Methods

- A commercially available LLM was used to generate 20 medically related questions using the prompt: "Please generate 20 medical questions like 'What is the prevalence of Sleep Apnea in the U.S.A?' on various medical topics".
- These questions were then submitted to two medical LLMs freely available to healthcare professionals.
- References cited in the generated answers were reviewed for indirect source referencing and claims not supported by the cited source (inaccurate claims).

Results

- Across 40 answers (20 per model), indirect source referencing was common, and only 12 (30%) answers used fully direct and accurate references.
- 26/40 (65%) included at least one **indirect** reference.
- 9/40 (22.5%) contained **inaccurate** claims (direct or indirect).
- 7/40 (17.5%) included **indirect** references with an **inaccurate** claim.
- 22/40 (55%) included at least one **direct** and **accurate** reference.
- The breakdown of reference directness and accuracy is shown in Figure 1 and Table 1.

Discussion

- Both medical LLMs frequently relied on indirect source referencing.
- Although one model was superior in achieving direct and accurate refencing, the percentages found in this study remain low.
- These findings suggest that, even when citations are provided, clinicians and editors should not assume primary-source verification by LLMs; manual checking of original studies remains necessary.
- Larger, multi-topic evaluations are warranted to assess generalizability, prompt designs, and system guardrails to reduce indirect referencing and citation inaccuracies.

Table 1. Citation directness and accuracy by large language models

Large Language Model Source	*Source 1*	*Source 2*
All direct references	9	3
All Direct w/inaccurate claim	0	2
At least one indirect reference	4	6
At least one Indirect reference w/ inaccurate claim	3	4
No direct references	4	5
Total	20	20

References

1. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):1-8.
2. Agarwal A, Arafa M, Avidor-Reiss T, Abdel-Meguid Hamoda TAA, Shah R. Citation Errors in Scientific Research and Publications: Causes, Consequences, and Remedies. World J Mens Health. 2023;41(3):461.
3. Pavlovic V, Weissgerber T, Stanisavljevic D, Pekmezovic T, Milicevic O, Lazovic JM, et al. How accurate are citations of frequently cited papers in biomedical literature? Clin Sci (Lond). 2021;135(5):671.
4. Homeier D, Adams M, Lynch T, Cognetti D. Inaccurate Citations Are Prevalent Within Orthopaedic Sports Medicine Literature. Arthrosc Sports Med Rehabil.2024;6(1):100873.

Only 30% of LLM answers used fully direct and accurate source referencing