A18

Haroun Haque M.S. Advisor(s): Afshin Razi M.D.

Co-author(s): Matthew C Johnson, Matthew L Magruder, Alex Hahn, Ameer Tabbaa, Katherine Coyner

Learning Alongside Language Learning Models: Accuracy of ChatGPT for Literature Citations in Spinal Surgery

Background: AI tools like ChatGPT are increasingly used in orthopedic research, offering potential as research aids by referencing open-access sources. However, concerns about the legitimacy of AI-generated citations, due to frequent "hallucinations," persist. This study assesses the accuracy of citations generated by ChatGPT v3.5 and v4.0 for spinal surgery, evaluating the responsiveness to varying prompt specificity.
Methods: In August 2024, three queries with increasing specificity were entered into ChatGPT v3.5 and v4.0, requesting outlines with 30 cited sources for seven spinal procedures: ACDF, PLIF, laminectomy, kyphoplasty, foraminotomy, disc replacement, and microdiscectomy. Citations were classified as existent or nonexistent, and a Chi-square analysis assessed differences in citation accuracy.
Results: A total of 420 citations were generated (210 for each language model). Nonexistent citations occurred in 27.1% of GPT-3.5 outputs and 44.3% of GPT-4.0 outputs (p=.007). Laminectomy had the highest rate of nonexistent citations (100% in GPT-4.0). ACDF showed a notable discrepancy, with GPT-4.0 producing three times as many nonexistent citations as GPT-3.5 (60% vs 20%; p=.0015). Hallucination rates for kyphoplasty, foraminotomy, disc replacement, and microdiscectomy were similar between the two models, but microdiscectomy showed a significant difference (23.3% vs 46.7%; p=.05). GPT-4.0 failed to generate any verifiable citations for laminectomy.
Conclusion: ChatGPT v3.5 and v4.0 vary in citation reliability, with GPT-4.0 showing a significantly higher hallucination rate. Nearly half of GPT-4.0's citations were nonexistent, and even valid sources were more often misattributed. Laminectomy was the most problematic, with GPT-4.0 failing to generate any verifiable sources. These findings emphasize the need for rigorous fact-checking, as ChatGPT-generated references are often unreliable, and should not be used without thorough human verification.