Comparative Analysis of ChatGPT-4o and Copilot on the American Society for Surgery of the Hand Self-Assessment Examination: Accuracy and Artificial Intelligence Hallucination

Benjamín Nieves-López, Mohammad saeed Kahrizi, Keith Aziz

INTRODUCTION: ChatGPT, developed by OpenAI, and Copilot from Microsoft, are large language models (LLMs) trained on extensive datasets to generate human-like responses and learn from previous conversations and encounters to improve their outputs. The implementation of artificial intelligence (AI) models in orthopedic surgery, including hand surgery, has gained popularity. However, one of the major challenges posed by AI-generated content is the AI hallucination phenomenon, in which the AI generates well-formatted but entirely fictional information. Despite the increasing integration of LLMs into clinical and educational contexts, their rates of reference fabrication within hand surgery remain unexplored. This study aimed to: 1) evaluate the accuracy rate of ChatGPT and Copilot in answering questions from the American Society for Surgery of the Hand (ASSH) Self-Assessment Examination (SAE); 2) assess the types of references cited in their responses; and 3) compare their rates of reference fabrication.

METHODS: ChatGPT-4o and Copilot AI models were independently prompted with 200 questions from the 2024 ASSH-SAE, 80 of which included clinical images, radiographs or both. AI performance was compared to the mean percentage of correct responses by hand surgeons who completed the 2024 ASSH-SAE. Following the answer generation, a secondary prompt was requested to provide the supporting references for each question. The types of references were classified into six categories: articles, educational websites, articles and books, books and educational websites, articles and educational websites, and all three source types combined (articles, educational websites, and books). Reference fabrication was assessed by verifying the accuracy of AI-generated references through Google search, PubMed, Google Scholar, and Crossref. A chi-square test was used to compare the categorical accuracy between ChatGPT-4o and Copilot. Independent t-tests were conducted to examine the fabrication rate. Statistical significance was set at a p-value≤0.05.

RESULTS: ChatGPT-4o answered 77.5% (155/200) of the questions correctly, compared to Copilot's 59.5% (119/200) and hand surgeons' real-world mean correctness rate of 76.6%. In text-only questions, ChatGPT-4o and Copilot obtained an accuracy rate of 78.3% (94/120) and 64.2% (77/120), respectively. ChatGPT-4o answered correctly 76.3% (61/80) of the image-based questions compared to Copilot's 52.5% (42/80). When stratified by types of images, ChatGPT-4o obtained an accuracy rate of 76.1% (35/46) in questions with clinical images, 69.2% (18/26) with radiographs, and 100% (8/8) in questions involving both. Copilot answered correctly 52.2% (24/46) of the questions with clinical image, 42.3% (11/26) with radiographs, and 87.5% (7/8) in questions with both. ChatGPT-4o cited references in all its responses while Copilot failed to provide references in 9% of the answers. ChatGPT-4o cited articles in 49.0% of the questions as its reference, in 0.5% were educational websites, 26.5% were articles and books, 0.0% were books and educational websites, 16.5% were articles and educational websites, and 7.5% were the combination of all three sources types. Copilot used articles as its reference for 45.0% of the questions, 5.0% were educational websites, 9.0% were articles and books, 2.5% were books and educational websites, 25.5% were articles and educational websites, and 3.5% were the combination of all three sources types. However, 97.5% of the references provided by ChatGPT-4o were fabricated compared to Copilot's 82.0% fabrication rate.

DISCUSSION AND CONCLUSION: ChatGPT-4o demonstrated the highest overall accuracy among the AI models and slightly surpassed the hand surgeons' real-world average performance. Both AI models underperformed on image-based questions compared to text-only ones. These findings raise concerns about the practicality of ChatGPT and Copilot in hand surgery, where imaging plays a crucial role. Due to unavailable hand surgeon data for image-based question accuracy, direct comparison in this subdomain was not conducted. ChatGPT-4o and Copilot predominantly cited articles, either alone or in combination with other sources such as books or educational websites. However, both AI models exhibited alarmingly high rates of reference fabrication, raising concerns about their reliability in clinical education and practice. These findings underscore the need for a robust and systematic AI reference-checking processes that do not facilitate the dissemination of misinformation in the medical field. As AI capabilities are refined and advanced, particularly in medical image interpretation and citation accuracy, further research is warranted before integration into clinical decision and medical education.

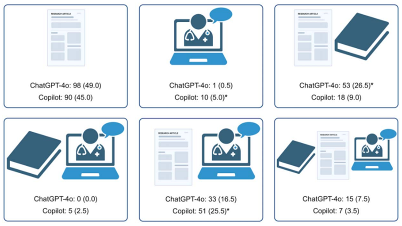

Table 1 ChatGPT-4o's performance compared to Copilot's

Question Type (total questions, n)	ChatGPT-4o n (%)	Copilot n (%)	P-value
Text-only (120)	94 (78.3)	77 (64.2)	0.02
Clinical Image (46)	35 (76.1)	24 (52.2)	0.03
Radiograph (26)	18 (69.2)	11 (42.3)	0.09
Clinical Image and Radiograph (8)	8 (100.0)	7 (87.5)	1.00

Table 2 AI Hallucination in ChatGPT-4o and Copilot

References	ChatGPT-4o n (%)	Copilot n (%)	P-value
No References	0 (0.0)	19 (9.5)	<0.01
Fabricated	195 (97.5)	164 (82.0)	0.63
Real	5 (2.6)	17 (8.5)	<0.01