Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253101

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties

Yasin Celal Güneş[1]
Turay Cesur[2]
Eren Çamur[3]

[1] Kırıkkale Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

[2] Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[3] Ankara [2]9 May State Hospital, Clinic of Radiology, Ankara, Türkiye

PURPOSE

This study aimed to compare six large language models (LLMs) [Chat Generative Pre-trained Transformer (ChatGPT)o1-preview, ChatGPT-4o, ChatGPT-4o with canvas, Google Gemini 1.5 Pro, Claude 3.5 Sonnet, and Claude 3 Opus] in generating radiology references, assessing accuracy, fabrication, and bibliographic completeness.

METHODS

In this cross-sectional observational study, 120 open-ended questions were administered across eight radiology subspecialties (neuroradiology, abdominal, musculoskeletal, thoracic, pediatric, cardiac, head and neck, and interventional radiology), with 15 questions per subspecialty. Each question prompted the LLMs to provide responses containing four references with in-text citations and complete bibliographic details (authors, title, journal, publication year/month, volume, issue, page numbers, and PubMed Identifier). References were verified using Medline, Google Scholar, the Directory of Open Access Journals, and web searches. Each bibliographic element was scored for correctness, and a composite final score [(FS): 0-36] was calculated by summing the correct elements and multiplying this by a 5-point verification score for content relevance. The FS values were then categorized into a 5-point Likert scale reference accuracy score (RAS: 0 = fabricated; 4 = fully accurate). Non-parametric tests (Kruskal–Wallis, Tamhane's T2, Wilcoxon signed-rank test with Bonferroni correction) were used for statistical comparisons.

RESULTS

Claude 3.5 Sonnet demonstrated the highest reference accuracy, with 80.8% fully accurate references (RAS 4) and a fabrication rate of 3.1%, significantly outperforming all other models (*P* < 0.001). Claude 3 Opus ranked second, achieving 59.6% fully accurate references and a fabrication rate of 18.3% (*P* < 0.001). ChatGPT-based models (ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview) exhibited moderate accuracy, with fabrication rates ranging from 27.7% to 52.9% and <8% fully accurate references. Google Gemini 1.5 Pro had the lowest performance, achieving only 2.7% fully accurate references and the highest fabrication rate of 60.6% (*P* < 0.001). Reference accuracy also varied by subspecialty, with neuroradiology and cardiac radiology outperforming pediatric and head and neck radiology.

CONCLUSION

Claude 3.5 Sonnet significantly outperformed all other models in generating verifiable radiology references, and Claude 3 Opus showed moderate performance. In contrast, ChatGPT models and Google Gemini 1.5 Pro delivered substantially lower accuracy with higher rates of fabricated references, highlighting current limitations in automated academic citation generation.

CLINICAL SIGNIFICANCE

The high accuracy of Claude 3.5 Sonnet can improve radiology literature reviews, research, and education with dependable references. The poor performance of other models, with high fabrication rates, risks misinformation in clinical and academic settings and highlights the need for refinement to ensure safe and effective use.

KEYWORDS

Reference, citation, ChatGPT o1-preview, Claude 3.5 Sonnet, large language models

Corresponding author: Yasin Celal Güneş

E-mail: gunesyasincelal@gmail.com

Received 28 October 2024; revision requested 10 December 2024; last revision received 21 February 2025; accepted 05 April 2025.

Epub: 12.05.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253101

You may cite this article as: Güneş YC, Cesur T, Çamur E. Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties. *Diagn Interv Radiol*. 2026;32(2):173-181.

The rapid advancement of large language models (LLMs) represents a key milestone in artificial intelligence (AI), offering unprecedented capabilities in text generation and comprehension.[1] These models, trained on extensive datasets, have shown promise in medical applications such as literature summarization, manuscript editing, and reference generation.[2,3] However, their reliability in reference generation remains a critical concern, particularly in radiology, where evidence-based practice depends on accurate and verifiable sources.[4,5] A key challenge is their tendency to generate "hallucinations" (fabricated or inaccurate references), which undermine their utility in clinical and academic settings.[5]

The issue of hallucinated references in LLMs is well documented in the literature.[6-16] Chelli et al.[7] reported hallucination rates of 39.6% for Chat Generative Pre-trained Transformer (ChatGPT)-3.5, 28.6% for ChatGPT-4, and an alarming 91.4% for Bard when generating references for systematic reviews. Walters and Wilder[8] found that although ChatGPT-4 exhibited a lower hallucination rate (18%) than ChatGPT-3.5 (55%), both models produced considerable inaccuracies, even among seemingly valid references. In radiology, Wagner et al.[9] observed that 63.8% of references generated by ChatGPT-3 were fabricated, with only 37.9% offering adequate support. These findings are particularly concerning in radiology, where inaccurate references could contribute to misinformation, potentially affecting clinical research, educational materials, and evidence-based decision-making.[9]

Main points

- Claude 3.5 Sonnet demonstrated the highest reference accuracy, significantly outperforming other large language models (LLMs) across all radiology subspecialties, making it the most reliable tool for generating medical references.
- Chat Generative Pre-trained Transformer (ChatGPT)-4o, ChatGPT-4o with canvas, and Google Gemini 1.5 Pro exhibited lower reference accuracy, with considerable inconsistencies in generating accurate references, highlighting the need for further improvements in these models for use in clinical settings.
- Accurate reference generation by Claude 3.5 Sonnet supports its potential to enhance literature reviews, research preparation, and educational content creation in radiology, improving the efficiency and quality of work in both clinical and academic domains.
- The study emphasizes the necessity of validating LLM-generated references, as errors and inconsistencies in models such as ChatGPT and Google Gemini could lead to serious risks in clinical decision-making and academic integrity.

Retrieval-augmented LLMs combine traditional language models with external data retrieval mechanisms, grounding responses in current, domain-specific information.[17] Emerging solutions, such as retrieval-augmented LLMs and platforms like OpenEvidence, aim to address these limitations by integrating real-time access to credible sources.[18] OpenEvidence, for instance, delivers up-to-date, evidence-based answers with clearly labeled references, reducing the risk of misinformation.[18] However, its accessibility remains restricted, requiring a National Provider Identifier number, which is issued to U.S. healthcare providers, for unlimited access and is available only in certain regions. In contrast, advanced LLMs such as ChatGPT-4o with canvas, ChatGPT o1-preview, and Claude 3.5 Sonnet offer worldwide accessibility, making them versatile and inclusive tools for users across diverse geographies.[19] These models have the potential to overcome prior limitations by leveraging enhanced natural language processing capabilities and expanded datasets, ensuring broader applicability and impact.[20]

Despite the rapid advancements in LLMs, no systematic evaluation has been conducted to assess the accuracy of references generated by state-of-the-art LLMs across radiology subspecialties. To address this gap, this study aims to provide the first systematic evaluation of the reference-generation accuracy of advanced LLMs, with a focus on identifying the most reliable model and characterizing variability across eight radiology subspecialties. By highlighting their strengths and limitations, this research seeks to clarify the potential roles of LLMs in radiology and provide actionable guidance for improving AI-driven reference generation.

Methods

Study design

This cross-sectional observational study evaluated the performance of six LLMs—ChatGPT o1-preview, ChatGPT-4o, ChatGPT-4o with canvas, Google Gemini 1.5 Pro, Claude 3.5 Sonnet, and Claude 3 Opus—in generating medical references for radiology questions across eight subspecialties. The study exclusively used publicly available, internet-based data without any identifiable patient information, eliminating the need for ethics committee approval. It was conducted in accordance with the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of LLMs in Healthcare guidelines.[21] An overview of the workflow is presented in Figure 1.

Question preparation

Eight radiology subspecialties—neuroradiology, abdominal imaging, musculoskeletal radiology, thoracic imaging, pediatric radiology, cardiac imaging, head and neck radiology, and interventional radiology—were selected to represent a broad range of clinical domains. For each subspecialty, 15 questions were developed, yielding a total of 120 questions. This sample size not only balances comprehensive coverage with the feasibility of manual reference verification but also exceeds the minimum requirement of approximately 96 questions—calculated using a standard sample size formula for estimating a 50% proportion with a 10% margin of error at the 95% confidence level—thus ensuring robust statistical power and enhancing the precision of our findings.

All questions were independently created by Radiologist 1 (Y.C.G.) without the use of any LLMs, thereby preventing any influence from the models' internal training data and minimizing potential bias from "leaked" context. All questions are provided in Supplementary Material 1.

Design of input–output procedures and performance evaluation for large language models

The input prompt was initiated as follows: "I am solving a radiology quiz and will provide you with open-ended, text-based questions. Please act as a radiology professor with 30 years of experience. Provide clear, comprehensive, and detailed answers to each question. Each answer must include four references to papers indexed in Medline. The references should include in-text citations as well as complete details, including the authors' names, title, journal, publication year, month, volume, issue, page numbers, and PubMed identifier (PMID)" (Figure 2). This prompt was presented in December 2024 on six distinct platforms with default parameters: OpenAI's ChatGPT o1-preview, ChatGPT-4o, ChatGPT-4o with canvas (https://chat.openai.com), Google Gemini 1.5 Pro (https://gemini.google.com), Claude 3.5 Sonnet, and Claude 3 Opus (https://claude.ai).

The allocation of tasks among the radiologists was as follows:

- Radiologist 2 (T.C.) conducted the questioning of ChatGPT-4o with canvas, Google Gemini 1.5 Pro, and ChatGPT o1-preview and recorded the responses.

- Radiologist 3 (E.Ç.) conducted the questioning of ChatGPT-4o, Claude 3.5 Sonnet, and Claude 3 Opus.

Due to resource limitations, the experiments were conducted with a single response per model per question to establish a standardized baseline. All LLMs were operated using their default parameters; only the first complete response generated by each model for each question was recorded. Notably, the LLMs were not pre-trained on any specific prompts, data, or question set prior to this study.

Reference evaluation

Validation of reference authenticity

Although the query requested Medline-indexed references, multiple databases were used for verification to account for possible indexing inconsistencies and to ensure a comprehensive assessment of reference accuracy. Each reference was verified across

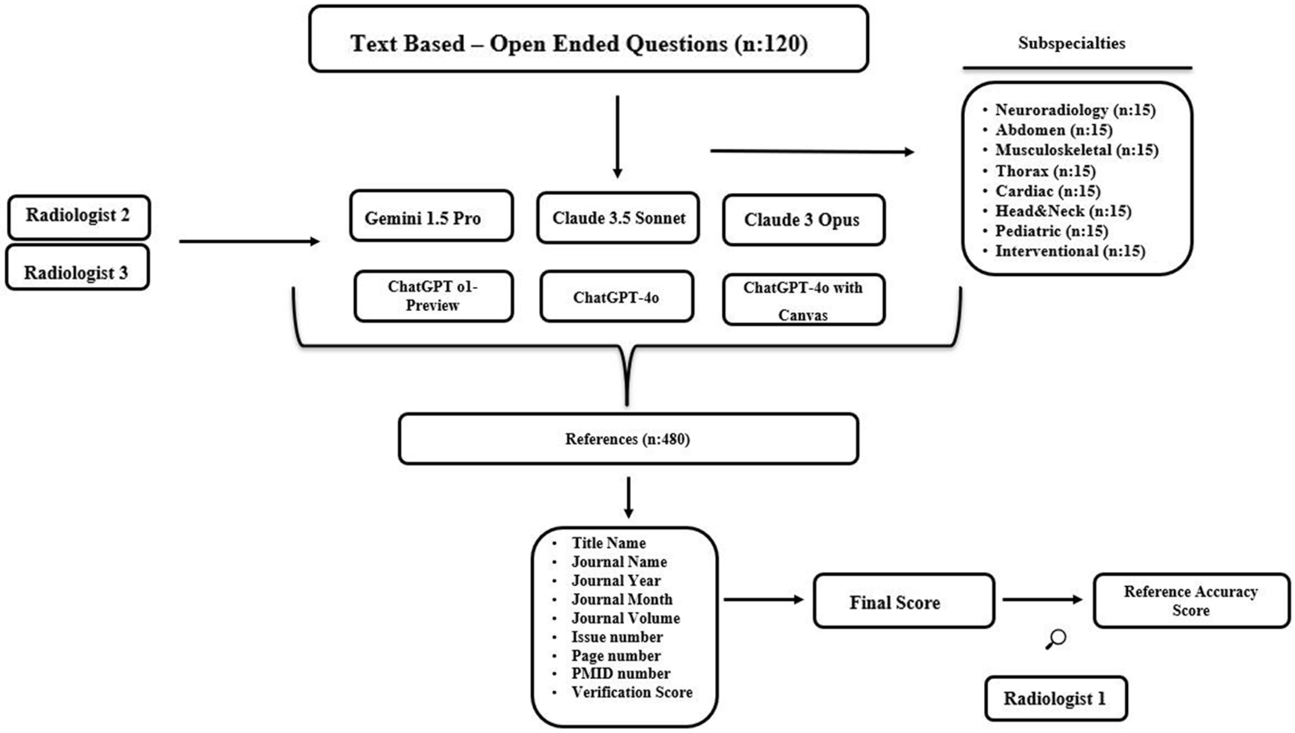
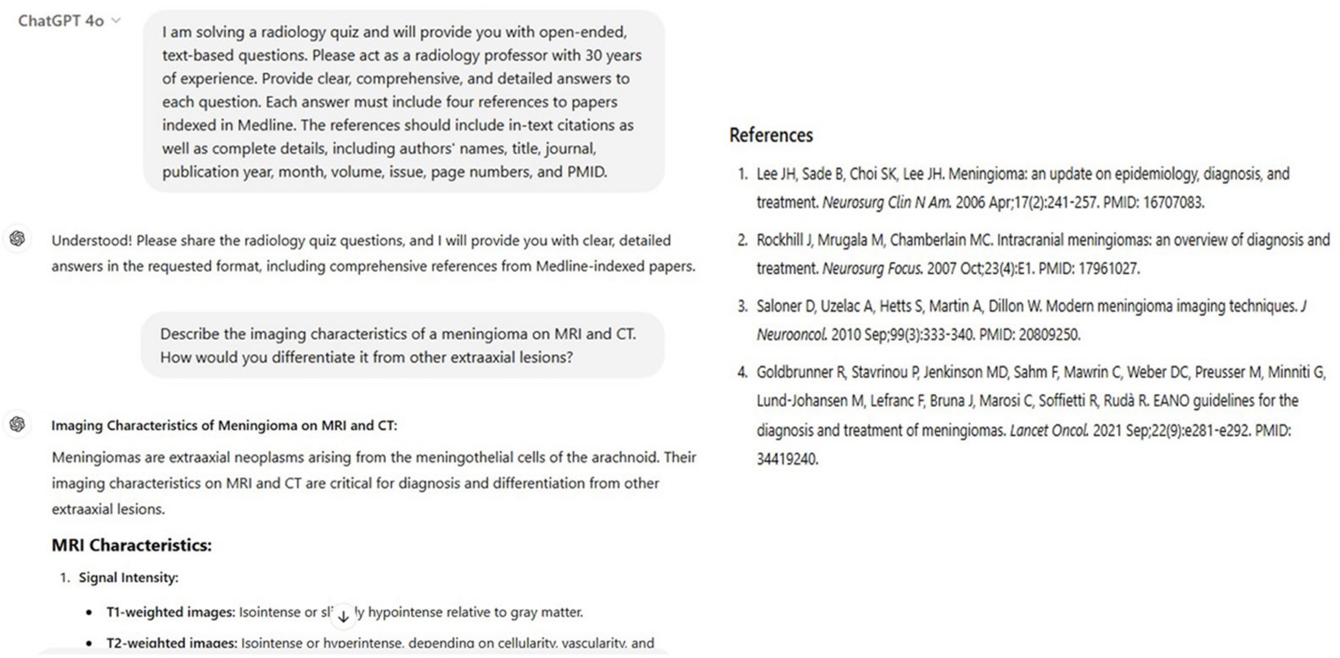


Figure 1. Overview of the study workflow.


ChatGPT 4o

I am solving a radiology quiz and will provide you with open-ended, text-based questions. Please act as a radiology professor with 30 years of experience. Provide clear, comprehensive, and detailed answers to each question. Each answer must include four references to papers indexed in Medline. The references should include in-text citations as well as complete details, including authors' names, title, journal, publication year, month, volume, issue, page numbers, and PMID.

Understood! Please share the radiology quiz questions, and I will provide you with clear, detailed answers in the requested format, including comprehensive references from Medline-indexed papers.

Describe the imaging characteristics of a meningioma on MRI and CT. How would you differentiate it from other extraaxial lesions?

Imaging Characteristics of Meningioma on MRI and CT:

Meningiomas are extraaxial neoplasms arising from the meningothelial cells of the arachnoid. Their imaging characteristics on MRI and CT are critical for diagnosis and differentiation from other extraaxial lesions.

MRI Characteristics:

1. **Signal Intensity:**
   - **T1-weighted images:** Isointense or sl ly hypointense relative to gray matter.
   - **T2-weighted images:** Isointense or hyperintense, depending on cellularity, vascularity, and

References

1. Lee JH, Sade B, Choi SK, Lee JH. Meningioma: an update on epidemiology, diagnosis, and treatment. *Neurosurg Clin N Am.* 2006 Apr;17(2):241-257. PMID: 16707083.
2. Rockhill J, Mrugala M, Chamberlain MC. Intracranial meningiomas: an overview of diagnosis and treatment. *Neurosurg Focus.* 2007 Oct;23(4):E1. PMID: 17961027.
3. Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. *J Neurooncol.* 2010 Sep;99(3):333-340. PMID: 20809250.
4. Goldbrunner R, Stavrinou P, Jenkinson MD, Sahm F, Mawrin C, Weber DC, Preusser M, Minniti G, Lund-Johansen M, Lefranc F, Bruna J, Marosi C, Soffietti R, Rudà R. EANO guidelines for the diagnosis and treatment of meningiomas. *Lancet Oncol.* 2021 Sep;22(9):e281-e292. PMID: 34419240.

Figure 2. Illustration of the prompts given to large language models and the corresponding responses they generated. MRI, magnetic resonance imaging; CT, computed tomography.

three databases—Medline, Google Scholar, and the Directory of Open Access Journals—and an internet search. If a reference could not be located in any of these databases, it was classified as fabricated.

Stylistic and bibliographic accuracy check

Although references were ultimately scored using a composite measure, each bibliographic element was explicitly examined:

- Authors' names (A), article title (T), journal name (J), publication year (Y), publication month (M), journal volume (V), issue number (I), page numbers (P), PMID number (PM).

Verification score

The verification score (VS) evaluates the accuracy and relevance of references generated by LLMs. Although LLMs may cite sources from the literature, it is crucial for authors to verify that the cited material precisely matches the phrase or statement being referenced. This ensures the accuracy and validity of the reference. To facilitate this evaluation, references are scored using a 5-point Likert scale:

- **0:** Reference is fabricated (not indexed).
- **1:** No pertinent information found in the source.
- **2:** Some pertinent information present.
- **3:** Largely pertinent information.
- **4:** Entirely pertinent information.

Reference accuracy score

The reference accuracy score (RAS) provides a unified metric for evaluating the bibliographic and verification accuracy of references. It is calculated using the following formula:

$$RAS = (A + T + J + Y + M + V + I + P + PM) \times VS$$

Each bibliographic element (A, T, etc.) is assigned 1 for a match or 0 for a mismatch. The VS, which reflects the alignment between the content and the cited source, is added to the total. This approach ensures a comprehensive evaluation, with scores ranging from 0 (fabricated) to 36 (fully accurate).

To facilitate interpretation, the RAS is categorized into a 5-point Likert scale:

- **RAS 0:** final score (FS) = 0 (fabricated)
- **RAS 1:** FS = 1–11 (weak accuracy)
- **RAS 2:** FS = 12–23 (moderate accuracy)
- **RAS 3:** FS = 24–35 (near accuracy)
- **RAS 4:** FS = 36 (fully accurate)

This categorization simplifies interpretation, offering a clear understanding of reference accuracy, from entirely fabricated to fully verified. Figure 3 provides a visual representation of the calculation and classification methods.

Radiologists' background

Three board-certified radiologists, each with 6 years of radiology experience, participated in this study. Radiologist 2 and radiologist 3 asked the questions to LLMs and recorded all answers. Radiologist 1 then evaluated all references and assessed the accuracy of the responses in a blinded manner, thereby minimizing the risk of bias.

Statistical analysis

Descriptive statistics, including medians, interquartile ranges (IQR), frequencies, and percentages, were calculated. The normality of variable distributions was assessed using the Kolmogorov–Smirnov test.

Due to the non-parametric distribution of the data, the Kruskal–Wallis test was employed to compare quantitative data across multiple groups (different LLMs). Following the Kruskal–Wallis test, Tamhane's T2 test was used for multiple post-hoc comparisons to identify specific group differences. Additionally, the Wilcoxon signed-rank test with a Bonferroni correction was applied to compare paired samples of RASs between LLMs. Statistical significance was set at $P < 0.003$ after applying the Bonferroni correction for 15 pairwise comparisons across six LLMs; otherwise, a P value < 0.05 was considered statistically significant. All statistical analyses were performed using SPSS version 28.0 (IBM Corp., Armonk, NY, USA).

Results

Reference accuracy by large language models

A total of 480 references were analyzed to compare the performance of the six LLMs. The evaluation focused on overall fabrication rates as well as stylistic and bibliographic accuracy across nine core components of each reference.

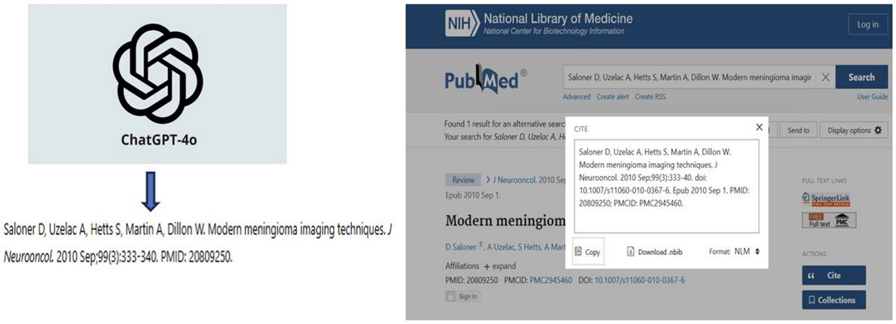
Stylistic and bibliographic accuracy

Authors' names and titles

Claude 3.5 Sonnet showed the highest accuracy for A (96.5%) and T (96.5%), followed by Claude 3 Opus at 81.7% for A and 81.3% for T. The ChatGPT-based models—ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview—generally fell in the mid-range, with accuracies between 44.8% and 58.5% for A and between 46.0% and 53.5% for T. Gemini 1.5 Pro performed the worst in both categories, reaching 38.5% for A and 40.2% for T.

Journal name, year and month

An analogous hierarchy appeared when evaluating J. Here, Claude 3.5 Sonnet again led at 95.6%, followed by Claude 3 Opus at 79.2%. The ChatGPT models ranged from 45.6% to 53.1%, and Gemini 1.5 Pro achieved 38.3%. For Y, Claude 3.5 Sonnet and Claude 3 Opus scored 95.6% and 77.7%, respec-



Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. *J Neurooncol.* 2010 Sep;99(3):333-340. PMID: 20809250.

Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imagir

Found 1 result for an alternative searc
Your search for *Saloner D, Uzelac A, H*

CITE

Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. J Neurooncol. 2010 Sep;99(3):333-40. doi: 10.1007/s11060-010-0367-6. Epub 2010 Sep 1. PMID: 20809250; PMCID: PMC2945460.

Copy Download .nbib Format: NLM

J Neurooncol. 2010 Se
Epub 2010 Sep 1.

Modern meningiom

D Saloner [1], A Uzelac, S Hetts, A Mart

Affiliations + expand

PMID: 20809250 PMCID: PMC2945460 DOI: 10.1007/s11060-010-0367-6

EVALUATION OF REFERENCE

(**Authors' Names (A)** – 1 Point + **Article Title (T)** – 1 Point + **Journal Name (J)** - 1 Point + **Publication Year (Y)** – 1 Point + **Publication Month (M)** - 1 Point + **Journal Volume (V)** – 1 Point + **Issue Number (I)** – 1 Point + **Page Numbers (P)** - 1 Point+ **PMID Number (PM)** - 1 Point) x **Verification Score (VS)** - 4 Point = 36 Point

Reference Accuracy Score (RAS) = 4 (Fully Accurate Reference) / Final Score 36

Figure 3. The example showcases the formatting of a reference generated by ChatGPT-4o, followed by its verification on PubMed. Each reference component, including author names, article title, journal name, publication year, month, volume, issue number, page numbers, and PMID, contributes to the final reference accuracy score. ChatGPT, Chat Generative Pre-trained Transformer; PMID, PubMed identifier.

tively, whereas the ChatGPT group landed between 41.9% and 53.1%. Gemini 1.5 Pro showed a low 26.7%. In M, Claude 3.5 Sonnet recorded 95.6% versus Claude 3 Opus at 77.3%, with the ChatGPT models coming in between 13.8% and 23.1% and Gemini 1.5 Pro at 31.7%.

Journal volume, issue number, and page number

Performance remained consistent for V, where Claude 3.5 Sonnet reached 95.2% and Claude 3 Opus 78.1%. The ChatGPT series ranged from 40.0% to 44.4%, and Gemini 1.5 Pro again dipped to 8.8%. Assessing I revealed 94.6% accuracy for Claude 3.5 Sonnet and 77.7% for Claude 3 Opus, with ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview spanning 29.8% to 42.7% and Gemini 1.5 Pro at 18.5%. For P, Claude 3.5 Sonnet and Claude 3 Opus recorded 93.8% and 77.5%, respectively, whereas ChatGPT-based models came in between 26.3% and 44.0%. Gemini 1.5 Pro once more ranked lowest at 16.5%.

PubMed identifier number

A similar pattern was seen in the PM category. Claude 3.5 Sonnet scored 94.0%, followed by Claude 3 Opus at 77.5%. The ChatGPT-4o model reached 23.1%, ChatGPT-4o with canvas 9.8%, ChatGPT o1-preview 10.8%, and Gemini 1.5 Pro was placed last at 3.3%.

Verification scores

VS showed a clear ranking among the LLMs. Claude 3.5 Sonnet and Claude 3 Opus both achieved the highest median verification Likert score of 4, with an IQR of 4–4 for each. In contrast, ChatGPT-4o recorded a median score of 3 (IQR: 0–4). ChatGPT-4o with canvas, ChatGPT o1-preview, and Gemini 1.5 Pro all had lower VSs, each reporting a median of 0 (IQR: 0–4).

Final scores of large language models

Final scores, presented as median and IQR, confirmed the leading positions of Claude 3.5 Sonnet and Claude 3 Opus. Claude 3.5 Sonnet ranked first with a median score of 36 (IQR: 36–36), followed by Claude 3 Opus at 36 (IQR: 36–18). ChatGPT o1-preview and ChatGPT-4o recorded median scores of 16 (IQR: 28–0) and 8 (IQR: 28–0), respectively. The lowest-ranked models were ChatGPT-4o with canvas with 0 (IQR: 28–0) and Gemini 1.5 Pro with 0 (IQR: 16–0).

All scores and reference component accuracies are summarized in Table 1.

Comparison of reference accuracy score by large language models

Claude 3.5 Sonnet exhibited the smallest fabrication rate at 3.1% while also achieving the highest proportion of fully accurate references (80.8%). Although Claude 3 Opus showed a higher fabrication rate of 18.3%, it still produced 59.6% fully accurate references. In comparison, the ChatGPT-based models all generated significantly more fabricated references (27.7%–52.9%) and fewer fully accurate ones (5.6%–7.3%). Gemini 1.5 Pro stood out with the highest fabrication rate of 60.6% and the lowest rate of fully accurate references at 2.7% (Table 2) (Figure 4).

Claude 3.5 Sonnet emerged as the top-performing model, significantly outperforming all others, including Claude 3 Opus ($P < 0.001$). Claude 3 Opus demonstrated strong performance, ranking second, with significant differences observed against all other models ($P < 0.001$). No significant differences were observed among the ChatGPT models. Specifically, comparisons of ChatGPT o1-preview and ChatGPT-o4 against ChatGPT-4o with canvas yielded Bonferroni-corrected *P* values of 0.019 and 0.037, respectively—both above the significance threshold of 0.003. Additionally, the difference between ChatGPT-4o and ChatGPT o1-preview was not significant ($P = 0.456$). In contrast, Google Gemini 1.5 Pro recorded the lowest accuracy, significantly underperforming compared with the Claude and ChatGPT models ($P < 0.001$) (Table 3).

Performance analysis by subspecialty

In a performance analysis of reference accuracy across multiple radiology subspecialties, several LLMs demonstrated distinct patterns of variability. Claude 3.5 Sonnet, Claude 3 Opus, ChatGPT-4o, ChatGPT o1-preview, and ChatGPT-4o with canvas each showed notable fluctuations ($P < 0.05$), whereas Google Gemini 1.5 Pro exhibited uniformly lower performance across all subspecialties without any statistically significant differences ($P > 0.05$) (Table 4).

Table 1. Comparative performance of large language models in reference component accuracy and overall scores

	Reference (n = 480)					
	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT-4o	ChatGPT o1-preview	ChatGPT-4o with canvas	Gemini 1.5 Pro
Authors' names	463 (96.5%)	392 (81.7%)	281 (58.5%)	251 (52.3%)	215 (44.8%)	185 (38.5%)
Title name	463 (96.5%)	390 (81.3%)	257 (53.5%)	250 (52.1%)	221 (46.0%)	193 (40.2%)
Journal name	459 (95.6%)	380 (79.2%)	219 (45.6%)	255 (53.1%)	220 (45.8%)	184 (38.3%)
Journal year	459 (95.6%)	373 (77.7%)	201 (41.9%)	248 (51.7%)	212 (44.2%)	128 (26.7%)
Journal month	459 (95.6%)	371 (77.3%)	66 (13.8%)	111 (23.1%)	68 (14.2%)	152 (31.7%)
Journal volume	457 (95.2%)	375 (78.1%)	204 (42.5%)	213 (44.4%)	192 (40.0%)	42 (8.8%)
Issue number	454 (94.6%)	373 (77.7%)	183 (38.1%)	205 (42.7%)	143 (29.8%)	89 (18.5%)
Page number	450 (93.8%)	372 (77.5%)	126 (26.3%)	211 (44.0%)	172 (35.8%)	79 (16.5%)
PMID number	451 (94.0%)	372 (77.5%)	111 (23.1%)	52 (10.8%)	47 (9.8%)	16 (3.3%)
Verification Likert score* [median, IQR (Q3-Q1)]	4 (4–4)	4 (4–2)	3 (4–0)	3 (4–0)	0 (4–0)	0 (4–0)
Final score [median, IQR (Q3-Q1)]**	36 (36–36)	36 (36–18)	8 (28–0)	16 (28–0)	16 (0–0)	0 (32–0)

IQR: interquartile range, Q1: 25% quantile, Q3: 75% quantile.

*Verification Likert score: this is categorized into a 5-point Likert scale reference accuracy score (0 = fabricated; 4 = fully accurate).

**Final score: the final score provides an integrated metric that combines the bibliographic accuracy of references with their verification score (VS). For each bibliographic element—such as authors' names, article title, journal name, and others—a match was scored as 1, and a mismatch was scored as 0. The VS, which measures how well the content aligns with the cited source, was then multiplied by the sum of the matched elements. PMID, PubMed identifier; ChatGPT, Chat Generative Pre-trained Transformer.

The post-hoc Tamhane test revealed that the Claude 3.5 Sonnet model showed no significant differences in reference accuracy across subspecialties, indicating uniformly consistent performance without any specific category demonstrating clear outperformance or underperformance. Similarly, Google Gemini 1.5 Pro performed uniformly across all subspecialties but with overall lower accuracy than other models.

Within Claude 3 Opus, neuroradiology demonstrated consistent superiority over most categories ($P < 0.05$), except for abdominal, cardiac, and head and neck radiology, where no significant differences were observed. Additionally, cardiac radiology outperformed the pediatric radiology group ($P =$ 0.020). No other significant differences were found among the remaining subgroups.

For ChatGPT-4o, cardiac radiology consistently emerged as the best-performing category ($P < 0.05$), except when compared with abdominal and interventional radiology, where performance was comparable. Conversely, pediatric radiology showed the weakest results, being significantly outperformed by other subspecialties, except for head and neck and musculoskeletal radiology ($P < 0.05$). No additional significant differences were detected.

In the case of ChatGPT-4o with canvas, thoracic radiology emerged as the highest-performing category, achieving significantly greater accuracy than most other subspecialties ($P < 0.05$), except for neuroradiology, cardiac, and musculoskeletal radiology. Conversely, head and neck radiology showed the weakest performance, being significantly outperformed by both thoracic radiology and cardiac radiology ($P < 0.05$). Additionally, cardiac radiology demonstrated superior performance to abdominal, pediatric, and interventional radiology ($P <$ 0.05). No further significant differences were observed among the subgroups.

As for ChatGPT o1-preview, head and neck radiology exhibited the lowest performance, being significantly outperformed by all other categories ($P < 0.05$) except for interventional and pediatric radiology, where no significant differences were observed. No further significant differences were identified among the subgroups.

Discussion

The most striking finding of our study is the consistent superiority of the Claude 3.5 Sonnet model in generating accurate and reliable medical references across diverse radiology subspecialties. With a significantly higher RAS ($P < 0.001$), a notably low fabrication rate (3.1%), and 80.8% of its references being fully accurate, Claude 3.5 Sonnet demonstrates a remarkable ability to integrate comprehensive radiological literature into its outputs. Given the critical importance of accuracy in reference generation, where even minor errors can have serious implications, Claude 3.5 Sonnet's ability to produce such a high percentage of fully accurate references underscores its potential as a reliable reference generator compared with other advanced LLMs. This superior performance likely stems from several factors, including a broader and more specialized training dataset and algorithmic refinements aimed at reducing hallucination rates—a common limitation in other models.[20] The Claude models leverage constitutional AI, a framework that prioritizes accuracy, ethical reasoning, and factual integrity, which may contribute to its minimized hallucination rates and enhanced reliability.[22]

In contrast, the Claude 3 Opus model, although ranking second overall, displayed a higher fabrication rate (18.3%) and a reduced proportion of fully accurate references (59.6%). This difference suggests that the underlying architecture of the Claude models is promising, especially in subspecialties where the training data may be less robust, such as pediatric or interventional radiology.

The ChatGPT models (ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview) exhibited only moderate performance. Their elevated rates of fabricated references—ranging from 27.7% to 52.9%—and recurrent inaccuracies in critical bibliographic components (such as PMID numbers and page details) indicate that these models have not yet achieved the precision required for reliable academic referencing. This result is consistent with prior studies on ChatGPT-generated medical content.[6-16] For instance, Bhattacharyya et al.[6] reported that nearly half the references produced by ChatGPT-3.5 were fabricated, with 47% being non-authentic and only 7% being both authentic and accurate. Similarly, Walters and Wilder[8] found that 55% of references from ChatGPT-3.5 were fabricated, and even in ChatGPT-4, the fabrication rate remained concerning at 18%, with 43% of authentic references from ChatGPT-3.5 and 24% from ChatGPT-4o containing substantive errors. Wagner et al.[9] evaluated ChatGPT-3's accuracy in answering 88 radiology questions and verifying references. Correct answers were provided for 67% of questions, and 33% contained errors. Of 343 references, 63.8% were fabricated, and only 37.9% of the verified references offered sufficient information.[9]

Gravel et al.[16] further observed that 69% of the 59 references generated by ChatGPT for medical questions were fabricated. In our study, ChatGPT-4o produced only 31 correct references out of 480, and ChatGPT o1-preview improved only modestly to 35 correct references, underscoring the persistent challenges in achieving accurate citation generation. These specific findings, along with the reported fabrication rates in our models, mirror the issues highlighted in the previous literature and indicate that even the upgraded versions of ChatGPT continue to fall short in reliably generating complete and verifiable academic references.

Table 2. Comparative evaluation of large language models based on reference accuracy score

	Reference (n = 480)					
RAS	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT-4o	ChatGPT-4o with canvas	ChatGPT o1-preview	Gemini 1.5 Pro
0 (fabrication)	15 (3.1%)	88 (18.3%)	133 (27.7%)	254 (52.9%)	226 (47.1%)	291 (60.6%)
1 (weak)	7 (1.5%)	18 (3.8%)	142 (29.6%)	22 (4.6%)	5 (1.0%)	21 (4.4%)
2 (moderate)	4 (0.8%)	21 (4.4%)	62 (12.9%)	32 (6.7%)	41 (8.5%)	74 (15.4%)
3 (near accurate)	66 (13.8%)	67 (14.0%)	112 (23.3%)	145 (30.2%)	173 (36.0%)	81 (16.9%)
4 (accurate)	388 (80.8%)	286 (59.6%)	31 (6.5%)	27 (5.6%)	35 (7.3%)	13 (2.7%)

Reference accuracy score: this evaluates the accuracy and relevance of references generated by large language models (LLMs). Although LLMs may cite sources from the literature, it is crucial for authors to verify that the cited material precisely matches the phrase or statement being referenced. This ensures the accuracy and validity of the reference. To facilitate this evaluation, references are scored using a 5-point Likert scale. ChatGPT, Chat Generative Pre-trained Transformer.

Google Gemini 1.5 Pro's performance was the poorest among the evaluated models, with a fabrication rate of 60.6% and only 2.7% of its references being fully accurate. The uniform underperformance of Google Gemini 1.5 Pro across all radiology subspecialties implies potential fundamental limitations—possibly stemming from a training dataset that underrepresents or insufficiently emphasizes medical literature or from an algorithmic framework that is less suited to the nuances of academic citation generation.

In our performance analysis by subspecialty, we highlighted that although Claude 3.5 Sonnet maintained uniformly high reference accuracy across all subspecialties, other models exhibited substantial variability. For example, Claude 3 Opus demonstrated superior performance in neuroradiology, whereas ChatGPT-4o achieved remarkable results in cardiac radiology and ChatGPT-4o with canvas showed exceptional performance in thoracic radiology. In contrast, Google Gemini 1.5 Pro consistently exhibited low accuracy across all subspecialties. These findings suggest that differences in data complexity and training representation may account for the inter-model and inter-subspecialty performance variations.

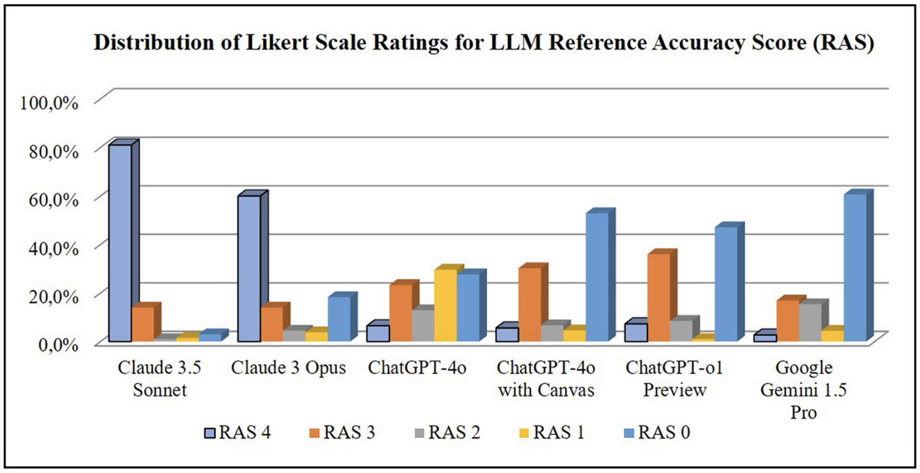


Figure 4. Distribution of Likert scale ratings for large language model reference accuracy scores. LLM, large language model.

Table 3. Comparison of accuracy of large language models with *P* values from the Wilcoxon test

	ChatGPT-4o	ChatGPT-4o with canvas	ChatGPT o1-preview	Google Gemini 1.5 Pro	Claude 3.5 Sonnet	Claude 3 Opus
ChatGPT-4o	-	0.037	0.456	<0.001	<0.001	<0.001
ChatGPT-4o with canvas	0.037	-	0.019	<0.001	<0.001	<0.001
ChatGPT o1-preview	0.456	0.019	-	<0.001	<0.001	<0.001
Google Gemini 1.5 Pro	<0.001	<0.001	<0.001	-	<0.001	<0.001
Claude 3.5 Sonnet	<0.001	<0.001	<0.001	<0.001	-	<0.001
Claude 3 Opus	<0.001	<0.001	<0.001	<0.001	<0.001	-

ChatGPT, Chat Generative Pre-trained Transformer.

Table 4. Reference accuracy score of large language models and classification by subspecialities

		Neuro	Abdomen	Musculoskeletal	Thorax	Cardiac	Head and Neck	Pediatric	Interventional	*P* value	
Claude 3 Opus	Median	4	4	4	3	4	4	3	4	<0.001	K*
	IQR (Q3–Q1)	(4–4)	(4–3.25)	(4–1)	(4–2)	(4–4)	(4–3)	(4–0)	(4–1)		
Claude 3.5 Sonnet	Median	4	4	4	4	4	4	4	4	0.008	K*
	IQR (Q3–Q1)	(4–4)	(4–4)	(4–4)	(4–3)	(4–4)	(4–3)	(4–4)	(4–3.25)		
ChatGPT-4o	Median	1	2	1	1	1	1	1	1	<0.001	K*
	IQR (Q3–Q1)	(2.75–0)	(3–0)	(2–0)	(3–1)	(3–1)	(3–0)	(1–0)	(3–1)		
ChatGPT-4o with canvas	Median	2	0	0	3	3	0	0	0	<0.001	K*
	IQR (Q3–Q1)	(3–0)	(3–0)	(3–0)	(3–0)	(3–0)	(2–0)	(2.75–0)	(2–0)		
ChatGPT o1-preview	Median	2	3	2.5	3	3	0	0.5	0	<0.001	K*
	IQR (Q3–Q1)	(3–0)	(3–0)	(3–0)	(3–0)	(3–0)	(1.5–0)	(3–0)	(3–0)		
Google Gemini 1.5 Pro	Median	0.5	0	0	0	0	0	0	0	0.244	K
	IQR (Q3–Q1)	(2.75–0)	(2–0)	(2–0)	(2–0)	(3–0)	(2–0)	(2–0)	(2.75–0)		

IQR, interquartile range; Q1, 25% quantile; Q3, 75% quantile; Neuro, neuroradiology; K, Kruskal–Wallis; *Tamhane's T2 test was used as a post-hoc comparison between each group, and the results of these comparisons are discussed in the results section. ChatGPT, Chat Generative Pre-trained Transformer.

Accurate reference generation is crucial in radiology, as evidence-based decision-making and scientific communication depend on verifiable and precise citations.[9] Inaccurate or fabricated references can lead to serious repercussions. For instance, misleading citations may result in clinicians basing diagnostic or treatment decisions on non-existent or irrelevant studies, ultimately affecting patient outcomes; in academic settings, reliance on erroneous citations can erode trust in literature reviews, undermine scholarly debates, and propagate errors in subsequent research.[23,24] Given these risks, the marked superiority of Claude 3.5 Sonnet has considerable practical implications, as this model could be integrated into workflows for manuscript preparation, automated literature retrieval, or even serve as an adjunct tool in clinical guideline development, provided that human experts continue to verify its outputs.

Additionally, our study observed that all the LLMs evaluated tend to favor references from the most well-known radiology papers. This tendency to prioritize widely cited papers can reinforce the "Matthew Effect," which refers to the phenomenon where frequently cited papers continue to gain references, overshadowing lesser-known but potentially important studies, in literature review processes.[25] This inclination of LLMs to rely on popular sources could narrow the scope of the literature being considered, limiting the diversity and range of research references. As a result, the use of these models may unintentionally contribute to reinforcing a limited set of references, reducing the overall richness of the academic discussion.

Although this study offers valuable insights into the capabilities of LLMs in generating medical references in radiology, several limitations must be noted. The dataset was relatively small, potentially limiting the generalizability of the findings across various radiological subspecialties and medical topics. Moreover, the use of a single standardized prompt may not capture the full variability of LLM responses arising from different prompting strategies or settings (e.g., temperature, top-K, top-P, and token limits). In addition, model performance was not assessed across multiple citation styles (e.g., AMA, Chicago), which restricts understanding of the broader applicability of these models in academic and clinical settings. The absence of repeated measurements for each LLM could introduce stochastic variability into the results, and the study evaluated only specific versions of LLMs available at the time, potentially misrepresenting the evolving capabilities of newer models. Future work may explore response consistency through multiple iterations per query.

In conclusion, Claude 3.5 Sonnet outperformed all other LLMs, demonstrating high accuracy and reliability in generating radiology references, making it well suited for tasks such as literature retrieval and manuscript preparation. This model holds great potential as a supportive tool for radiologic reference generation, offering a valuable resource to complement evidence-based practice. In contrast, other models exhibited higher fabrication rates and inconsistent accuracy, underscoring the need for substantial improvements. Future efforts should focus on enhancing performance in underperforming subspecialties and refining bibliographic accuracy to meet the rigorous demands of evidence-based radiology.

Acknowledgements

The authors thank Büşra Özal Güneş for his assistance with the statistical analysis of this manuscript. They utilized ChatGPT-4o, a large language model (January 2025 Version; OpenAI; https://chat.openai.com), to revise the grammar and English translation. The content of this publication is entirely the responsibility of the authors, who have reviewed and edited it as necessary.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary: https://d2v96fxpocvxx.cloudfront.net/c1dc3a38-51db-436b-af33-1bc7522029b3/content-images/a6628944-e199-444c-8b76-3c2a723824b2.pdf

References

1. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol.* 2024;34(5):2817-2825. [Crossref]
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]
3. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc*. 2023;16:1513-1520. [Crossref]
4. Goktas P, Agildere AM. Transforming radiology with artificial intelligence visual chatbot: a balanced perspective. *J Am Coll Radiol*. 2024;21(2):224-225. [Crossref]
5. Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: literature so far. *Curr Probl Diagn Radiol*. 2024;53(2):215-225. [Crossref]
6. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15(5):e39238. [Crossref]
7. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164. [Crossref]
8. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. [Crossref]
9. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J*. 2024;75(1):69-73. [Crossref]
10. Athaluri SA, Manthena SV, Kesapragada VSRKM, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432 [Crossref]
11. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024;281(4):2159-2165. [Crossref]
12. Day T. A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *Prof Geogr*. 2023;75(6):1024-1027. [Crossref]
13. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. 2023;326:115334. [Crossref]
14. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol*. 2023;280(11):5129-5133. [Crossref]
15. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J Med Internet Res*. 2024;26:e52935. [Crossref]
16. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin Proc Digit Health*. 2023;1(3):226-234. [Crossref]
17. Steybe D, Poxleitner P, Aljohani S, et al. Evaluation of a context-aware chatbot

using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. *J Craniomaxillofac Surg*. 2025;53(4):355-360. **[Crossref]**

18. Patel N, Grewal H, Buddhavarapu V, Dhillon G. OpenEvidence: enhancing medical student clinical rotations with AI but with limitations. *Cureus*. 2025;17(1): e76867. **[Crossref]**

19. Temsah MH, Jamal A, Alhasan K, Temsah AA, Malki KH. OpenAI o1-preview vs. ChatGPT in healthcare: a new frontier in medical AI reasoning. *Cureus*. 2024;16(10):e70640. **[Crossref]**

20. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in radiology's "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(12):1399-1402. **[Crossref]**

21. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865-868. **[Crossref]**

22. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *APJESS*. 2023;11(3):118-134. **[Crossref]**

23. Dumas-Mallet E, Boraud T, Gonon F. Le mésusage des citations et ses conséquences en médecine [Citation misuse and its effects on public health]. *Med Sci (Paris*). 2021;37(11):1035-1041. **[Crossref]**

24. Peoples N, Østbye T, Yan LL. Burden of proof: combating inaccurate citation in biomedical literature. *BMJ*. 2023;383:e076441. **[Crossref]**

25. Larivière V, Gingras Y. 2010. The impact factor's Matthew Effect: a natural experiment in bibliometrics. J Assoc Information Sci Technol 2010;61(2):424-427. **[Crossref]**