DATABASE ANALYSIS

e-ISSN 1643-3750
© Med Sci Monit, 2026; 32: e953782
DOI: 10.12659/MSM.953782

Received: 2026.04.16
Accepted: 2026.06.10
Available online: 2026.06.24
Published: 2026.XX.XX

Accuracy and Error Patterns of References Generated by Large Language Models in Endodontics: The Role of Prompt Design and Model Selection

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ABCDEF 1 Mehmet Adıgüzel
BDEF 2 Alparslan Mustafa Çeler

1 Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye
2 Department of Pediatric Dentistry, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye

Corresponding Author: Mehmet Adıgüzel, Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Tayfur Sökmen Campus, 31060, Alahan, Antakya, Hatay, Türkiye, Phone: +903262456060, e-mail: dt.mehmetadiguzel@gmail.com

Financial support: None declared
Conflict of interest: None declared

Background: Large language models (LLMs) are increasingly used in healthcare; concerns persist regarding the accuracy of generated bibliographic references. The effect of prompt design on reference reliability has not been clearly established. This comparative experimental study evaluated the impact of prompt specificity on LLM-generated reference accuracy in endodontics and compared model performance.

Material/Methods: We used ChatGPT 5 and Claude Sonnet 4.6. Ten predefined endodontic queries were combined with 3 prompt types of increasing specificity. Each model generated 5 references per query-prompt combination (total: 300 references). References were verified using PubMed, Google Scholar, and CrossRef. Accuracy was classified as fabricated (0), partially accurate (1; existing references containing ≥1 bibliographic inaccuracy), or fully accurate (2). Digital object identifier (DOI) accuracy was assessed separately. Statistical analyses were performed using mixed-effects models and Pearson's chi-square test or Fisher's exact test.

Results: Accuracy scores tended to increase with greater prompt specificity ($P = 0.249$). Claude demonstrated significantly higher accuracy than ChatGPT (mean score: 1.79 vs 1.25; $P < 0.001$). DOI accuracy did not differ among prompt groups ($P = 0.338$); it was significantly higher for Claude than for ChatGPT (90.0% vs 35.3%; $P < 0.001$). ChatGPT produced significantly more title, journal, and DOI errors ($P < 0.001$); author and year errors were similar between models.

Conclusions: Prompt specificity had limited effects on reference accuracy; model selection played a greater role. DOI accuracy was strongly model-dependent and largely unaffected by prompt design under the test conditions, highlighting the need for external verification of LLM-generated references.

Keywords: Endodontics • Artificial Intelligence • Bibliography • Data Accuracy • Reproducibility of Results

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/953782

3243 5 — 22

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Introduction

The rapid development of large language models (LLMs), including systems such as ChatGPT and Claude, has reshaped access to scientific information in healthcare and dental research [1-3]. These models are increasingly used for tasks such as literature retrieval, evidence synthesis, clinical decision support, and education [4-7]. In endodontics, recent evidence suggests that LLMs can generate clinically relevant responses for diagnostic scenarios, treatment planning, and the management of procedural complications, highlighting their potential as adjunctive tools in both research and clinical practice [8,9]. Endodontics was selected as the focus of this study because it is a highly evidence-based dental specialty that frequently relies on precise bibliographic referencing and rapidly evolving scientific literature.

Despite these promising applications, concerns persist regarding the reliability and factual accuracy of LLM-generated outputs. A well-recognized limitation is the phenomenon of "hallucination," in which models produce information that appears plausible but lacks verifiable evidence [10-12]. In academic contexts, this issue is particularly relevant when LLMs generate bibliographic references that are partially inaccurate or entirely fabricated. Previous studies in medical and dental fields have shown that LLM-generated citations may include incorrect authors, mismatched journal names, inaccurate publication details, and invalid digital object identifiers (DOIs), raising concerns about their suitability for scientific use [12-14].

Accurate referencing is essential to scientific integrity, particularly in the context of emerging artificial intelligence (AI) applications [13]. Citations play central roles in supporting claims, maintaining transparency, and enabling reproducible research. Inaccurate or unverifiable references can undermine the credibility of scientific work and contribute to the spread of misleading information, especially in specialized fields such as endodontics, where clinical decisions often depend on high-quality evidence. In endodontics, inaccurate or unverifiable references may affect the interpretation of scientific evidence and hinder evidence-based clinical decision-making. Therefore, bibliographic accuracy is a matter of academic reporting and an important prerequisite for the responsible use of AI-generated information in clinical practice.

Recent investigations have shown that inaccuracies in LLM-generated references may arise both from completely fabricated citations and from "partial hallucinations," in which correct bibliographic elements are combined with incorrect or inconsistent details [15]. Current evidence also suggests that different LLMs exhibit distinct error patterns; some models are more prone to metadata-related inaccuracies than others [16].

Whereas existing studies have primarily focused on overall accuracy and hallucination patterns, the factors influencing bibliographic reliability remain insufficiently explored. In particular, the role of prompt design—an essential component of human-AI interaction—has not been systematically evaluated in the context of reference generation. Given that LLM outputs are highly sensitive to input phrasing, variations in prompt specificity might influence the accuracy and structure of generated references [17]. Such potential influence has contributed to the emergence of prompt engineering as a strategy to improve output quality; however, its ability to ensure bibliographic accuracy remains unclear, particularly in endodontics.

Therefore, the aim of the present study was to evaluate the effect of prompt specificity on the accuracy of bibliographic references generated by LLMs in endodontics and to compare the performance of 2 contemporary models.

This study was designed to test the following prespecified hypotheses: (1) increasing prompt specificity from Group A to Group C would improve overall reference accuracy scores; (2) increasing prompt specificity would improve DOI accuracy; and (3) model selection would significantly affect both overall accuracy and DOI accuracy. The primary outcome was predefined as overall reference accuracy score, whereas DOI accuracy was considered a secondary outcome.

Material and Methods

Study Design

This study was designed to evaluate the effect of prompt specificity on the accuracy of LLM-generated bibliographic references. A comparative experimental design was used in which different prompt structures were applied to the same predefined set of endodontic queries.

Models Evaluated

ChatGPT (GPT-5, OpenAI, San Francisco, CA, USA) and Claude (Sonnet 4.6, Anthropic, San Francisco, CA, USA) were evaluated through their publicly available web interfaces in April 2026. No additional system-level configurations or parameter adjustments were applied.

Ethical Considerations

Ethical approval was not required for this study because it did not involve human participants, animal subjects, or patient data.

Query Selection

In total, 10 predefined endodontic queries were developed to represent clinically relevant and commonly investigated topics

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

in endodontics. The queries covered a range of subjects, including postoperative pain, irrigation, working length determination, canal morphology, and treatment outcomes.

All queries were prepared in English and remained unchanged across all experimental conditions. The full list of queries and prompt structures is provided in the Supplementary Material. The selected queries were intended to represent common clinical and research topics in endodontics, ensuring both thematic diversity and clinical relevance.

Prompt Design

Prompts were categorized into 3 groups based on their level of instructional specificity:
Group A (Simple Prompt):
A general instruction requesting references without additional constraints.
Group B (Constraint-Based Prompt):
A prompt specifying that the references should be indexed in PubMed.
Group C (Accuracy-Focused Persona Prompt):
A prompt instructing the model to act as an expert academic librarian and generate only real, verifiable references with correct DOI information, without including fabricated citations.

All prompts were predefined and applied identically across both models to ensure consistency. These prompt categories were intended to reflect increasing levels of instructional specificity and to simulate common real-world user interactions, ranging from general queries to more constrained and accuracy-focused requests.

Reference Generation

For each query, all 3 prompt types were independently submitted to each model. Each prompt was entered in a separate, independent session to avoid contextual influence from previous interactions.

For every query-prompt combination, the models were instructed to generate 5 references. In total, 300 references were generated (10 queries × 3 prompt groups × 2 models × 5 references). A single-response approach was adopted to simulate a real-world usage scenario in which users typically rely on a single model output per query rather than performing repeated sampling.

Reference Verification

Each generated reference was independently verified across multiple bibliographic databases, including PubMed, Google Scholar, and CrossRef. PubMed served as the primary database because of its widespread use in clinical research and its established role as a standard indexing source for biomedical literature.

The following bibliographic components were assessed: existence of the reference, author names, article title, journal name, publication year, and DOI accuracy. Author-list truncation using "et al." was not considered an error when the first author or initial authors were correctly reported and the reference corresponded to a verifiable article. However, incorrect or inconsistent author information was classified as an error. All references were independently evaluated by 2 investigators, and discrepancies were resolved by consensus.

Scoring System

Each reference was assigned an accuracy score as follows: score 0 (fabricated; the reference was not found in any database), score 1 (partially accurate; the reference existed but contained ≥ 1 bibliographic inaccuracy), and score 2 (fully accurate; all bibliographic details were correct). DOI accuracy was recorded separately as correct or incorrect. A reference was classified as fabricated only if no corresponding record could be identified in any of the consulted databases.

Error Classification

Errors were categorized as fabricated references, incorrect DOIs, author mismatches, title inaccuracies, journal mismatches, and year inconsistencies. These categories were selected because they represent the principal bibliographic elements required for accurate identification, retrieval, and verification of scientific publications; they have often been evaluated in previous studies concerning the accuracy of LLM-generated citations.

Statistical Analysis

No formal power analysis was performed because the study design was based on a predefined and balanced experimental structure intended to ensure standardized comparisons across models, prompt groups, and query categories. The number of generated references was determined by the experimental design rather than participant recruitment; all predefined query-prompt-model combinations were included in the analysis. Statistical analyses were performed at the output level to address clustering of references within each query-prompt-model combination and to enable standardized comparisons of mean accuracy and DOI accuracy across models and prompt groups. Mixed-effects models were used, whereby model, prompt group, and their interaction comprised fixed effects and query served as a random effect. Mean accuracy scores and DOI accuracy rates were analyzed as dependent variables. Comparisons of categorical error frequencies between groups were performed using Pearson's

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 1. Mean accuracy scores and DOI accuracy rates according to model and prompt group (output-level analysis).

Model	Prompt group	Mean accuracy score, mean ± SD	DOI accuracy, mean ± SD
ChatGPT	A	1.22 ± 0.26	0.36 ± 0.22
ChatGPT	B	1.24 ± 0.21	0.30 ± 0.22
ChatGPT	C	1.28 ± 0.21	0.40 ± 0.25
Claude	A	1.70 ± 0.22	0.86 ± 0.13
Claude	B	1.78 ± 0.26	0.90 ± 0.17
Claude	C	1.88 ± 0.19	0.96 ± 0.08

Note: Values are presented as mean ± standard deviation. Each value represents an output-level estimate aggregated across queries (unique combination of model, prompt group, and query; n = 10 outputs per cell). DOI accuracy values represent output-level mean proportions (eg, 0.86 = 86% DOI accuracy). Abbreviations: DOI, digital object identifier; SD, standard deviation.

Table 2. Mixed-effects analysis of mean accuracy and DOI accuracy at the output level.

Outcome	Effect	F	df	*P* value
Mean accuracy	Model	85.95	1, 54	< 0.001
	Prompt group	1.43	2, 54	0.249
	Model × Prompt group	0.35	2, 54	0.704
DOI accuracy	Model	133.91	1, 54	< 0.001
	Prompt group	1.11	2, 54	0.338
	Model × Prompt group	0.37	2, 54	0.693

Note: Mixed-effects models were used to evaluate the effects of model, prompt group, and their interaction on output-level mean accuracy and DOI accuracy. Abbreviation: DOI, digital object identifier.

chi-square test or Fisher's exact test, as appropriate, based on expected cell counts. Odds ratios (ORs) with 95% confidence intervals (CIs) were also calculated for model-based error comparisons. *P* values < 0.05 were considered statistically significant.

Results

Overall, 60 outputs (2 models × 3 prompt groups × 10 queries) and 300 references were included in the analysis. Output-level analyses were conducted to address clustering of references within each query-prompt-model combination.

Output-Level Analysis

The descriptive results are presented in **Table 1**. Across all prompt groups, Claude consistently exhibited higher mean accuracy scores and DOI accuracy rates than ChatGPT (mean accuracy score: 1.79 vs 1.25, *P* < 0.001; DOI accuracy: 90.0% vs 35.3%, *P* < 0.001).

Mixed-effects analysis demonstrated a significant main effect of model on mean accuracy score ($F(1,54) = 85.95$, $P < 0.001$), indicating that overall reference accuracy significantly differed between models. In contrast, prompt group was not statistically significant ($F(2,54) = 1.43$, $P = 0.249$). No significant model-by-prompt interaction was observed ($F(2,54) = 0.35$, $P = 0.704$), indicating that the effect of prompt specificity did not vary according to model (**Table 2**).

A similar pattern was observed for DOI accuracy. Model selection had a significant effect ($F(1,54) = 133.91$, $P < 0.001$), whereas prompt group did not ($F(2,54) = 1.11$, $P = 0.338$). The interaction between model and prompt group also was not significant ($F(2,54) = 0.37$, $P = 0.693$) (**Table 2**). This finding may be relevant when LLM-generated references are used for literature retrieval and verification.

Reference-Level Accuracy

At the reference level, Claude produced a substantially higher proportion of fully accurate references compared with ChatGPT. Specifically, 79.3% of Claude-generated references

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 3. Distribution of reference-level accuracy scores by model.

Model	Score 0, n (%)	Score 1, n (%)	Score 2, n (%)	Total, n
ChatGPT	1 (0.7)	111 (74.0)	38 (25.3)	150
Claude	1 (0.7)	30 (20.0)	119 (79.3)	150

Accuracy score: 0 = fabricated; 1 = partially accurate; 2 = fully accurate.

Table 4. Comparison of error types between models.

Error type	ChatGPT, n (%)	Claude, n (%)	OR (95% CI)	*P* value
Title error	70 (46.7)	8 (5.3)	15.63 (7.09-34.48)	< 0.001
Journal error	40 (26.7)	6 (4.0)	8.70 (3.57-21.28)	< 0.001
DOI error	94 (62.7)	14 (9.3)	16.39 (8.55-31.25)	< 0.001
Author error	18 (12.0)	19 (12.7)	0.94 (0.47-1.87)	1.000
Year error	4 (2.7)	2 (1.3)	2.03 (0.37-11.24)	0.680

Note: ORs represent the odds of error occurrence in ChatGPT relative to Claude. P values were calculated using Fisher's exact test.
Abbreviations: CI, confidence interval; DOI, digital object identifier; OR, odds ratio.

Table 5. Prompt-group comparisons of reference error types.

Error type	Group A, n/N (%)	Group B, n/N (%)	Group C, n/N (%)	Test	P value	Cramer's V
DOI error	37/100 (37%)	39/100 (39%)	32/100 (32%)	χ^2	0.569	0.061
Title error	24/100 (24%)	37/100 (37%)	17/100 (17%)	χ^2	0.005	0.189
Journal error	13/100 (13%)	24/100 (24%)	9/100 (9%)	χ^2	0.010	0.176
Author error	18/100 (18%)	6/100 (6%)	13/100 (13%)	χ^2	0.035	0.150
Year error	3/100 (3%)	2/100 (2%)	1/100 (1%)	Fisher's exact	0.600	0.058

Abbreviation: DOI, digital object identifier.

were fully accurate (score = 2), whereas only 25.3% of ChatGPT-generated references met this criterion. In contrast, most ChatGPT-generated references were classified as partially accurate (score = 1), representing 74.0% of all references (**Table 3**).

Error Pattern Analysis

Error pattern analysis showed that DOI-related inaccuracies were the most frequent type of error, followed by title- and journal-related errors. These errors were significantly more common in ChatGPT outputs than in Claude outputs ($P < 0.001$ for all comparisons).

Specifically, DOI errors were observed in 62.7% of ChatGPT-generated references, compared with 9.3% of Claude-generated references. Title errors were identified in 46.7% of ChatGPT outputs and 5.3% of Claude outputs, whereas journal-related errors occurred in 26.7% and 4.0% of references, respectively.

In contrast, no statistically significant differences were found between models for author-related errors ($P = 1.000$) or year-related errors ($P = 0.680$), indicating that these types of inaccuracies were relatively infrequent and comparable across models (**Table 4**). The largest effect sizes were observed for DOI- and title-related errors. Compared with Claude-generated references, ChatGPT-generated references showed substantially higher odds of DOI errors (OR = 16.39, 95% CI: 8.55-31.25), title errors (OR = 15.63, 95% CI: 7.09-34.48), and journal errors (OR = 8.70, 95% CI: 3.57-21.28) (**Table 4**).

Prompt-group analyses demonstrated significant differences in title-, journal-, and author-related errors, whereas DOI- and year-related errors did not significantly differ among prompt groups (**Table 5**). According to Cramer's V values, the observed effect sizes were small to modest.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Discussion

The present study evaluated the effect of prompt specificity on the accuracy of bibliographic references generated by LLMs and compared the performance of 2 contemporary models. The findings partially supported the predefined hypotheses. Although a gradual increase in accuracy scores was observed with increasing prompt specificity, this effect was not statistically significant. In contrast, model selection had significant impacts on overall accuracy and DOI accuracy, suggesting that model-dependent factors play a more decisive role than prompt design.

Previous studies have shown that LLM-generated references can contain hallucinations, metadata-related inaccuracies, and model-dependent differences. However, the effect of prompt specificity on bibliographic reliability has not been systematically evaluated within a controlled, domain-specific context. In this context, the present study provides a focused and incremental contribution by examining reference accuracy in endodontics, distinguishing between overall accuracy and DOI accuracy, and demonstrating that prompt specificity influences certain metadata-related errors—but not DOI accuracy—under the test conditions. Additionally, the single-response design reflects a real-world usage scenario and enables more practical interpretation of model performance.

It has been noted that only a limited proportion of LLM-generated references are fully accurate, with high rates of fabricated and partially inaccurate citations. For example, Bhattacharyya et al [18] demonstrated that a minority of references generated by ChatGPT were entirely correct, underscoring the risk of bibliographic inaccuracies in AI-generated content. Within endodontics, the effect of prompt specificity on bibliographic reference accuracy has not been specifically examined in a structured comparative framework. The present study extends the existing literature by providing a domain-specific analysis and distinguishing between overall accuracy vs DOI accuracy. The findings indicate that prompt specificity does not significantly influence DOI reliability, although it affects certain metadata-related errors. The present results imply that model selection, rather than prompt refinement alone, is a more decisive factor in determining bibliographic reliability.

Accuracy scores gradually increased with greater prompt specificity; however, this pattern was not statistically significant. Previous research has also shown that increases in prompt specificity do not necessarily improve citation accuracy. Linardon et al [19] reported that more specific or constrained prompts can even increase the likelihood of fabricated or inaccurate references, particularly in specialized topics. These findings suggest that prompt engineering alone is insufficient to ensure bibliographic reliability, given that its effect on accuracy appears limited despite widespread use. Model outputs may be more strongly influenced by underlying architecture and training data than by prompt refinement alone [17].

In contrast, the models substantially differed. Claude demonstrated better performance across all major outcome measures, including overall accuracy and DOI accuracy. A difference in DOI accuracy is particularly important because the DOI serves as a key identifier for verifying and retrieving scientific literature. Previous cross-disciplinary studies have also focused on the DOI as a particularly error-prone component in LLM-generated citations, with a high prevalence of incorrect or nonfunctional identifiers [20]. These findings further emphasize the need for external verification of DOI information in LLM-generated references and highlight the importance of model selection when LLMs are used in academic and clinical contexts [2,21].

Analysis of error patterns provides further insight into the nature of inaccuracies in LLM-generated references. ChatGPT showed significantly higher rates of title, journal, and DOI errors, suggesting a greater tendency toward distortion of bibliographic metadata rather than complete fabrication. In contrast, author and year errors were relatively infrequent and comparable between models. These findings support the distinction between fabricated references and partially inaccurate references; they suggest that most errors arise from incorrect reconstruction of bibliographic details, rather than from the generation of entirely nonexistent articles. This distinction may also explain the relatively low number of fully fabricated references observed in the present study. Because complete hallucinations and partial hallucinations were classified separately, the strict definition of fabrication may have reduced the observed frequency of fully hallucinated outputs. Other classification approaches may yield higher estimates of complete hallucination rates. Additionally, some LLM instances, particularly those involving ChatGPT, tended to shorten author lists using “et al.”; such cases were not classified as errors when the cited article was verifiable and the reported author information was otherwise consistent.

Prompt specificity influenced certain metadata-related errors, including title, journal, and author inaccuracies. However, DOI-related errors were largely unaffected by prompt design. This finding suggests that DOI generation depends on deeper internal mechanisms within LLMs that are less responsive to surface-level prompt modifications. From a practical perspective, this finding indicates that even carefully designed prompts may be insufficient to ensure reliable DOI output.

From a clinical perspective, these findings are directly relevant to endodontic research and clinical decision-making, where accurate referencing remains essential for evidence-based

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

practice and the interpretation of emerging scientific literature. Recent studies have shown that LLMs can provide clinically relevant responses in endodontic diagnosis, treatment planning, patient communication, and decision-support applications [8-10]. However, the present findings suggest that bibliographic foundations supporting such outputs are not consistently reliable. Ozbay et al [8] reported that LLMs may assist clinical decision-making in endodontics, whereas Demir Cicek and Cicek [9] highlighted the potential roles of LLMs in patient education regarding root canal treatment. Notably, Suarez et al [10] demonstrated that although ChatGPT showed high response consistency, its accuracy remained limited. The present findings extend these observations by revealing that inaccuracies in generated references might represent an additional challenge when LLM outputs are used to support evidence-based endodontic practice. Although LLMs can facilitate rapid insights, their outputs require careful verification before use. In particular, DOI information should not be considered accurate without external validation. These findings suggest that clinicians and researchers should exercise caution when using LLM-generated references, particularly with respect to DOI-based verification. Inaccurate references may contribute to misinformation, thus affecting both scientific reporting and evidence-based clinical practice.

Previous medical and dental studies have shown that LLM-generated references may appear plausible while containing substantial bibliographic inaccuracies [12,14]. The present findings extend this concern to endodontics and suggest that DOI accuracy is among the most vulnerable components of LLM-generated citations.

Several limitations of this analysis should be considered. First, each model was queried only once per topic; only a single response was obtained for each query-prompt combination. Because LLMs are probabilistic, repeated queries can yield different outputs. Accordingly, the present findings should be assumed to reflect single-instance interactions, rather than overall model consistency. Second, the analysis was limited to a predefined set of endodontic queries, and the findings may not be generalizable to other domains. Third, formal a priori power analysis was not performed because the study utilized a fixed, balanced design based on predefined query-prompt-model combinations, rather than conventional participant recruitment. Consequently, the study may have had limited ability to detect smaller effect sizes, and the findings should be interpreted within this methodological context. Fourth, only 2 LLMs were evaluated, which may limit the generalizability of the findings to other available models. Fifth, formal inter-rater reliability statistics were not calculated, although all references were independently evaluated by 2 investigators and discrepancies were resolved through consensus. Therefore, the consistency of evaluations between investigators could not be formally quantified. Finally, model performance may change over time due to system updates, which could affect reproducibility [22].

Despite these limitations, the present study provides a structured and reproducible framework for evaluating the bibliographic accuracy of LLM-generated references. Future research may build on this approach by including additional models, repeated-query designs, or alternative prompt strategies, as well as by examining domain-specific differences in performance. These findings support the careful and critically informed use of LLMs in scientific writing.

Conclusions

Within the limitations of this study, prompt specificity had limited effects on the accuracy of LLM-generated references; model selection appeared to play a more decisive role in ensuring reliable bibliographic outputs under the test conditions. Claude demonstrated higher overall accuracy and DOI reliability than ChatGPT. These findings should be interpreted within the specific context of the experimental conditions, predefined endodontic queries, the 3 prompt types evaluated, the 2 models tested, and the single-response workflow. The effects of prompt design may vary with repeated sampling, alternative prompting strategies, different domains, or updated model versions. Careful manual verification of LLM-generated references remains essential before use in academic or clinical contexts.

Acknowledgments

AI-assisted language-editing tools were used during manuscript preparation. All outputs were critically reviewed and verified by the authors.

Department and Institution Where Work Was Done

Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye.

Patient Consent

Not applicable.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

References:

1. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: A comparative analysis. BMC Oral Health. 2025;25(1):573
2. Umer F, Batool I, Naved N. Innovation and application of large language models (LLMs) in dentistry: A scoping review. BDJ Open. 2024;10(1):90
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. J Dent Res. 2020;99(7):769-74
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56
5. Surlari Z, Budală DG, Lupu CI, et al. Current progress and challenges of using artificial intelligence in clinical dentistry: A narrative review. J Clin Med. 2023;12(23):7378
6. Alharbi SS, Alhasson HF. Exploring the applications of artificial intelligence in dental image detection: A systematic review. Diagnostics (Basel). 2024;14(21):2442
7. Pastucha M, Skarżyński H, Kochanek K, Jedrzejczak WW. Reference accuracy in large language model chatbots: A metric for inherent misinformation? Med Sci Monit. 2026;32:e950916
8. Ozbay Y, Erdogan D, Dincer GA. Evaluation of the performance of large language models in clinical decision-making in endodontics. BMC Oral Health. 2025;25(1):648
9. Demir Cicek B, Cicek O. Evaluating the response of AI-based large language models to common patient concerns about endodontic root canal treatment: A comparative performance analysis. J Clin Med. 2025;14(21):7482
10. Suarez A, Diaz-Flores Garcia V, Algar J, et al. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-13
11. Giannakopoulos K, Kavadella A, Aaqel Salim A, et al. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. J Med Internet Res. 2023;25:e51580
12. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus. 2023;15(2):e35179
13. Schwendicke F, Sidhu SK, Ferracane JL, et al. Generative AI: Opportunities, risks, and responsibilities for oral sciences. J Dent Res. 2025;104(13):1429-31
14. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):14045
15. 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
16. 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
17. Hassanein FE, Ahmed Y, Maher S, et al. Prompt-dependent performance of multimodal AI model in oral diagnosis: A comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. Sci Rep. 2025;15(1):37932
18. 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
19. Linardon J, Jarman HK, McClure Z, et al. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. JMIR Ment Health. 2025;12:e80371
20. 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
21. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198
22. Van Dis EA, Bollen J, Zuidema W, et al. ChatGPT: Five priorities for research. Nature. 2023;614(7947):224-26

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]