



Published in final edited form as:

Laryngoscope. 2026 March ; 136(3): 1109–1117. doi:10.1002/lary.70165.

Generative Artificial Intelligence Methodology Reporting in Otolaryngology: A Scoping Review

Isaac L. Alter, AB¹, Karly Chan¹, Katerina Andreadis, MS², Anaïs Rameau, MD, MSc, MPhil, MS¹

¹Sean Parker Institute for the Voice, Department of Otolaryngology-Head and Neck Surgery, Weill Cornell Medical College, New York, NY

²Department of Population Health, New York University Grossman School of Medicine, New York, NY

Abstract

Objective: Researchers in otolaryngology-head and neck surgery (OHNS) have sought to explore the potential of large language models (LLMs), but many publications do not include crucial information, such as prompting approach and model parameters. This has substantial implications for reproducibility, since LLMs can generate different output based on differences in “prompt engineering.” We aimed to critically review methodological reporting and quality of LLM-focused literature in OHNS.

Data Sources: Databases were searched in October 2024, including PubMed, Embase, Web of Science, ISCA Archive, IEEE Xplore, arXiv, medRxiv, and engRxiv.

Review Methods: Abstract and full text review, as well as data extraction, were performed by two independent reviewers. All primary studies using LLMs within OHNS were included.

Results: From 925 abstracts retrieved, 117 were included. All studies used ChatGPT, with a minority (16.2%) including additional LLMs. Only 46.2% published direct quotations of all prompts. While the majority (76.9%) reported the number of prompts, only 6.8% rationalized this number, while 23.9% reported the number of runs per prompt. Most publications (73.5%) provided some description of prompt development, though only 11.1% explicitly described why specific decisions in prompt design were made, and only 6.0% reported prompt testing. There was no evidence that quality of methodology reporting was improving over time.

Conclusion: LLM-focused literature in OHNS, while exploring many potentially fruitful avenues, demonstrates variable completeness in methodological reporting. This severely limits the generalizability of these studies and suggests that best practices could be further disseminated and enforced by researchers and journals.

CORRESPONDING AUTHOR: Anaïs Rameau, MD MPhil MS FACS, Sean Parker Institute for the Voice, Department of Otolaryngology – Head and Neck Surgery, Weill Cornell Medical College, 240 E 59th St, New York, NY, 10022, Anr2783@med.cornell.edu.

CONFLICTS OF INTEREST: None

A portion of this research was previously presented at the American Head and Neck Society Meeting from May 14–16 in New Orleans, LA.

Keywords

Artificial intelligence; large language models; natural language processing; otolaryngology

INTRODUCTION

Since the release of ChatGPT to the public in 2022, interest in artificial intelligence (AI), machine learning (ML), and specifically large language models (LLMs) has skyrocketed. The healthcare sector has been no exception, with countless use cases being devised and theorized.^{1–4} Within otolaryngology-head and neck surgery (OHNS), investigations into the efficacy and utility of LLMs have likewise proliferated rapidly, with reports focused on diagnosis, clinical decision making, writing skills, and test-taking abilities, to name just a few.^{5,6}

However, additional scholarship has delineated that different LLMs, and even seemingly similar queries to a given LLM, are not equivalent. Firstly, LLMs can be fine-tuned – i.e. additional training data can be incorporated – to conform to a specific task. LLM responses can also be affected by subtle changes in the prompt given to the model; this has given rise to “prompt engineering,” or the design of prompts to optimize desired output.⁷ Many studies have extensively documented the wide variability in LLMs’ ability to answer biomedical and clinical questions based on different prompting strategies.^{8–12} A recent investigation additionally recommended running prompts multiple times, as this can uncover inconsistencies in LLM responses that one-time prompting will often miss.¹³

What follows from the reliance of LLMs on prompting is a potential lack of generalizability of LLM-focused studies: two investigations assessing similar clinical abilities of the same LLM could find markedly different results based solely on their prompting strategies. There are also implications for reproducibility – the ability of individuals outside the original research team to independently arrive at similar findings with the same methodology and data.¹⁴ With the recent introduction of reasoning models, prompting recommendations between models now vary significantly.^{15,16} This necessitates rigorous, transparent and standardized reporting of methodology in studies querying LLMs. Unfortunately, these practices are not uniform in LLM-focused research at large.⁷ This leads to potential misperceptions of LLMs’ abilities, which, when implications regarding healthcare are raised, can have significant ethical repercussions.¹⁷ Additional ethical concerns – such as patient privacy and algorithmic bias – also necessitate careful approaches within research using LLMs, though these concepts are often not represented in existing publications.⁴

These factors have led to the development of consensus statements recommending reporting criteria, such as TRIPOD-LLM and the DEAL-B checklist.^{18,19} These criteria include not only a complete description of the LLM being used, but also details about prompt development and fine-tuning, evaluation of output, use of patient-specific data, and discussion of ethical issues such as bias. However, while overall LLM use within OHNS research has been well-catalogued by recent scoping reviews, these publications have not assessed methodology or methodological reporting.^{5,6}

Given the known profound impacts of different methodological approaches to LLMs, as well as the uneven compliance with recommendations regarding descriptions of methodology in biomedical research at large, we sought to undertake a scoping review of LLM-focused literature within OHNS, to assess quality and consistency of methodology reporting and to evaluate reproducibility of the findings within LLM research in our specialty.

METHODS

This review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISM-ScR) checklist.²⁰ The review protocol was delineated in advance and made available on October 7, 2024 on Open Science Framework (https://osf.io/pzyec/?view_only=e55078e6f3414545be942ad872adfd7c).

Search Strategy

A search strategy was devised in collaboration with a scholarly medical librarian, combining keywords encompassing both OHNS and LLMs. Searches were run on October 19, 2024, in Pubmed, Embase, Web of Science, ISCA Archive, and IEEE Xplore, as well as gray literature sources including engrXiv, arXiv, and medrXiv. Full search strategies are available in Appendix 1. Search results were limited to publications released after November 2022, when ChatGPT first became publicly available.

In order to reflect more recently published literature in this field, identical search terms were used to search for publications within June of 2025.

Study Selection

All primary studies within OHNS using LLMs were included, regardless of application or methodology. Studies without a full text available in English were excluded. In addition, non-primary literature such as reviews and commentaries were excluded. In the format of a PICOS statement, we included studies involving any participants (or no participants), examining the use of LLMs, with no necessary comparison, evaluating any chosen outcome regarding the LLM's performance, and with no restriction on study design. Search results were collected in EndNote reference manager and subsequently uploaded to Covidence screening software; duplicates were automatically removed. All abstracts were screened independently by two reviewers (I.L.A. and K.C.). Full texts of eligible studies were independently reviewed by the same researchers. Disagreements between reviewers were resolved by consensus, or by consultation with the senior author (A.R.). Inter-rater reliability for both the abstract and full text phases was calculated via Cohen's Kappa.

Data Extraction

A data extraction template was created by the entire research team, and was tested iteratively to ensure consistent data capture. The variables extracted and their definitions are listed in Appendix 2. Extraction was carried out by two independent reviewers (I.L.A. and K.C.) in Covidence, with discrepancies resolved via consensus; data was subsequently exported

to Microsoft Excel. Descriptive statistics were calculated to determine what proportion of included studies reported particular methodological details.

RESULTS

After removal of duplicates, 925 unique studies were captured by the literature search. Full texts were reviewed for 157 publications; 40 were ultimately excluded, yielding 117 studies deemed appropriate for inclusion. Cohen's Kappa for the abstract and full text screening were 0.84 and 0.78, respectively, corresponding to "near-perfect" and "substantial" agreement. Study selection is summarized via a PRISMA flow diagram in Figure 1. Included studies are listed in Supplemental Table 1, and full citations are included in Appendix 3. Recent reviews have already robustly catalogued the prevalence and primary applications of LLMs within OHNS and other fields, thus this basic information for our review's included studies is included in Appendix 4 and additionally summarized in Supplemental Figure 1 and Supplemental Figure 2.^{6,21,22} The results of the additional search for publications from June 2025 are summarized in Appendix 5 and Supplemental Table 2 and compiled as citations in Appendix 6; the results reported herein reflect only the initial search.

Reporting on LLM Name, Version, and Parameters

All of the 117 included studies used a form of ChatGPT. Nineteen publications (16.2%) compared GPT with other LLMs such as Google's Gemini (previously known as Bard) (15/117, 12.8%), Claude (3/117, 2.6%), or Meta's Llama (3/117, 2.6%). Other LLMs appearing in one publication each were Mistral, Bing (now renamed Copilot), WebMD Symptom Checker, Perplexity AI, PaLM2, and MedPaLM. Eighteen studies (18.4%) compared different versions of ChatGPT, such as GPT-3.5 and GPT-4.0; of note, three used GPT-4 with vision (GPT-4V), which includes a "computer vision" component to analyze image inputs as well as text. These results are summarized in Figure 2A. Most studies used only one LLM (86/117, 73.5%), while 21 (17.9%) compared two LLMs. Six publications (5.1%) compared three LLMs, and two studies (1.8%) each compared four and five LLMs, respectively (Figure 2B).

Only 54 publications (46.2%) reported the date of LLM use, while 46 (39.3%) reported their method of accessing the LLM (i.e. via website, custom Application Programming Interface (API), etc.). Three studies (2.6%) used custom APIs. Four publications (3.4%) reported adjusting the baseline parameters of the LLM (such as temperature), while an additional six (5.1%) reported making no such alterations; the remaining 107 (91.5%) did not specify whether they were using default settings or making changes to LLM parameters. Three studies (2.6%) incorporated external data for fine-tuning, while two (1.7%) provided external data to the LLM via retrieval-augmented generation; all others presumably assessed LLMs' existing "knowledge," based on the training data already incorporated in their creation, though only six studies (5.1%) made specific mention of having not performed fine-tuning. Zero studies reported the number of tokens used by their LLM queries.

Prompt Methodology Reporting

Fewer than half of all studies (54/117, 46.2%) published direct quotations of all prompts; an additional 31 (26.5%) quoted one or a subset of prompts, while 32 (27.4%) only summarized the content of their prompts (Figure 3A). The majority of studies (90/117, 76.9%) reported the number of prompts used, with a mean (SD) of 148.2 (574.1) per study; number of prompts ranged from one to 4,666, with the largest subset of studies (66/117, 56.4%) using fewer than 50 prompts. The distribution of studies by number of prompts used is shown in Figure 3B. However, only eight studies (6.8%) provided a rationale for the number of prompts used, and among these, only two performed a power calculation to plan the number of prompts in advance.

Most publications (86/117, 73.5%) provided some description of how prompts were developed. However, the specificity of these descriptions was limited: only 13 studies (11.1%) gave explicit justifications of why specific decisions in prompt design were made (e.g. written from a patient perspective to assess symptom checker diagnostic accuracy). Fifty-six studies (47.8%) provided partial justification for such decisions, while 48 (41.0%) made no mention of their specific decision-making in prompt design. Strikingly, only seven studies (6.0%) reported using prompt testing or refinement, and only 28 (23.9%) explicitly reported the number of times each prompt was run.

Nine studies directly reported the number of accounts (7.7%) used to query the LLM(s), while 51 (43.6%) specified whether prompts were entered sequentially in the same chat window versus in separate chats.

Reporting of Data Evaluation

Eighty-two studies (70.1%) used no validated tool to assess LLM responses. Of the publications using validated scales, the most commonly employed (10/35, 28.6%) was the Patient Education Material Assessment Tool (PEMAT),²³ followed by DISCERN (7/35, 20%),²⁴ and the Artificial Intelligence Performance Instrument (AIPI) (4/35, 11.4%).²⁵ Most studies (71/117, 60.6%) reported the qualifications of the assessor of LLM output, and the majority of these (61/71, 85.9%) used board-certified otolaryngologists, sometimes in conjunction with other physicians, residents, or medical students. While 30 publications (25.6%) did not specify the number of assessors, among the remaining studies, four (4.6%) used one assessor, 37 (42.5%) used two, and 46 (52.9%) used more than two. 40 publications (34.2%) conducted an interrater reliability assessment.

Governance Considerations Reporting

Sixteen publications (13.6%) discussed the level of transparency or explainability of LLMs, while 40 (34.2%) acknowledged other potential ethical or governance issues surrounding LLM use, such as algorithmic bias or patient privacy. However, only one study actively tested LLMs for bias. Twenty-eight studies (23.9%) used patient data, though five of these pulled data from public sources such as social media posts including Reddit, Yahoo Answers, or Facebook. Of the remaining 23, 14 (60.8%) reported having obtained informed consent from participants; all of the studies not reporting informed consent took place outside the United States, largely from Europe. No included studies used facial or vocal

patient identifiers, and none mentioned strategies for using LLMs in a way that would allow for safe use of protected health information (such as integration with an electronic health record). Many studies reported anonymizing patient data, but did not specify what potentially identifying data was removed prior to LLM input.

DISCUSSION

In this scoping review, we have catalogued the methodology used in LLM research within OHNS literature and aimed to describe the quality, consistency and transparency of reporting to assess reproducibility. To the authors' knowledge, this manuscript represents the first scoping review to assess LLM methodology within OHNS. As the use of LLMs has skyrocketed both in the world at large and in OHNS research, a description of the current state of scholarship in this area is ever more important to ensure quality of research. Furthermore, since LLMs have introduced a previously untapped – and rapidly evolving – avenue for research, many studies have been published in the relative absence of standardized reporting guidelines, necessitating a thorough understanding of the quality of these investigations. Our review found marked variability in completeness of methodological reporting, with the vast majority of studies omitting key factors, especially with regards to description of LLM prompting strategies.

For example, fewer than half of the reviewed publications reported the date of LLM use; this is notable because LLM performance on identical tasks can vary over time.²⁶ In addition, very few studies reported adjusting model parameters or employing fine-tuning techniques to improve LLMs' performance. Because fine-tuning can have significant impacts on LLMs' abilities to answer medical questions or accomplish tasks^{27–29} – especially in a field such as OHNS with highly specific knowledge – not only is the use of fine-tuning more likely to maximize clinical applicability, but also the absence of specifying whether fine-tuning has taken place leads to significant questions about the generalizability of a study's results.

Relatedly, although most studies reported the number of prompts used, fewer than half published direct quotations of all prompts. Furthermore, a very small minority provided any rationale for the number of prompts used, such as performing a power calculation or otherwise justifying sample size, explicitly describing resource constraints limiting the number of prompts, or reporting reaching a point of saturation. Perhaps most strikingly, however, was that almost 90% of studies failed to provide a specific rationale for why prompts were written a particular way, and only 6% performed any kind of prompt testing or refinement. It is difficult to overstate the importance of prompting strategies in determining LLM outputs; some studies suggest that prompt engineering alone can be more productive than fine-tuning (which can often be time-consuming and/or expensive) for medical applications of LLMs.^{11,30} Even aspects as seemingly trivial as formatting can affect prompt outputs, necessitating the sharing of direct quotes or screenshots of prompts to ensure reproducibility.³¹ In addition, one publication found that careful attention to prompt design and iteration could reduce major LLM hallucinations by 75%.³² Thus, in the absence of precise reporting around the prompts used, the ways in which they were entered, and the decisions surrounding their development, the generalizability of studies employing LLMs within OHNS is thrown into question.

It is important to note that there is no evidence that the quality of reporting is improving over time; for example, of the 86 of our included studies published by June 2024, 43 (50%) included direct quotations of all prompts, compared with 11/31 (35.4%) of studies published since July 2024. The emergence of TRIPOD-LLM and DEAL-B as standardized frameworks for reporting LLM studies not only highlights the importance of publishing this information, but also provides easy-to-implement checklists for investigators carrying out such studies, potentially laying the groundwork for improved reporting in the future, since all included studies from the initial search were published prior to these formal reporting recommendations.^{18,19} However, the repeated search gathering studies published in June 2025 failed to show significant improvement; fewer than half of these studies provided direct quotations of all prompts, and only small minorities justified prompt design decisions, performed prompt testing or refinement, or rationalized the number of prompts used.

Several key concepts underpin most prompt engineering strategies. The first is the use of zero-shot, one-shot, or few-shot prompts – namely, giving the LLM zero, one, or multiple examples of an ideal output in the prompt.¹¹ Chain-of-thought prompting is another technique which specifically asks the LLM to perform tasks in a step-by-step fashion or explain its reasoning, while a related strategy known as heuristic prompting includes pre-defined rules in the prompt to aid the LLM in decision-making.⁸ Some comparisons of prompting strategies have found that zero-shot prompting was more likely to be successful for basic tasks, while unfamiliar questions or inputs requiring domain-specific knowledge often required few-shot prompts to maximize efficacy.^{9,10} However, other studies have suggested that optimal prompt strategies vary significantly between different LLMs and specific clinical applications.⁸ A potential difficulty with few-shot prompting is the requirement for additional tokens, or units of data used to process text; not only does this increase cost for paid versions of LLMs, but some studies also suggest that increased token counts are associated with worse performance.^{10,11}

Additional “best-practice” prompt design tools include asking questions as specifically as possible, directly describing the LLM’s role (also known as the “Persona Pattern,”), and testing and refining prompts until desired outputs are achieved.³³ It is important to note that beyond prompt engineering, additional factors – such as utilizing the LLM’s “memory” by asking questions asked in the same session, or adjusting model parameters such as temperature (a measure of randomness or “creativity”) – can also significantly influence LLMs’ outputs.^{33,34}

It is also important to note that relatively few studies explicitly discussed governance concerns surrounding LLMs, such as algorithmic bias and patient privacy. This is particularly important given that 65% of U.S. hospitals use AI and predictive models, though only 44% evaluate these models for bias.³⁵ In addition, while only a minority of reviewed publications entered patient data into LLMs, some of these studies did not report obtaining informed consent from individuals whose data was used. While evaluating LLMs’ abilities to provide accurate patient-facing information within OHNS has its value, LLMs can also make real provider-facing contributions to clinical workflows, such as extraction and summarization of electronic health record data,³⁶ writing notes,³⁷ and providing clinical decision-making support.³⁸ However, robust studies have demonstrated persistent biases

against minoritized or disadvantaged groups (such as Black, unhoused or LGBTQIA+ individuals) when LLMs are asked to provide healthcare recommendations.³⁹ It is thus crucial that investigators of LLMs within OHNS carefully consider concerns around ethics and privacy in order to advance our field and leverage the capabilities of LLMs to better serve patients.

Despite the important contributions of this review, there are relevant limitations to discuss. Firstly, given the nature of this manuscript as a scoping review, analysis of the results of the included studies is not within our purview, and we can thus not make any summative claims about the overall efficacy of LLMs in achieving the tasks set by OHNS researchers. In addition, certain classifications in our data extraction cannot be universally applied to all LLM studies; for example, LLM pipelines that use chain-of-thought techniques cannot be easily described in terms of the number of runs per prompt. Furthermore, as previously noted, the rate of publication – and of change – within the field of LLM research is incredibly rapid; thus, recently published studies from November 2024 to the present are not included in this review, limiting our ability to comment on potential improvements over time in methodological reporting quality. Importantly, despite the fact that prior standards for reporting in AI research predate the vast majority of LLM-focused OHNS research,⁴⁰ the most up-to-date recommendations regarding methodology reporting in LLM research – such as the aforementioned TRIPOD-LLM¹⁸ and DEAL-B¹⁹ – have been released after the publication of the studies included in the initial search for this review, and thus it could be supposed that the quality of reporting in our field has improved in recent months. We sought to address these limitations by performing an additional search cataloguing publications from June 2025, but we acknowledge that this provides only a snapshot of potential evolution in manuscript reporting. It is also important to note that despite basic comparisons to assess potential changes in quality of reporting over time, our methodology does not allow for more in-depth or definitive longitudinal analysis, and future studies could seek to more robustly address whether methodology reporting is improving in our field.

It is the authors' hope that illuminating the variability and shortcomings of methodology reporting in LLM research in our field can serve as a call to action for investigators and journals alike to strive for consistent, uniform, and reliable reporting in future studies, allowing for a higher-quality and more clearly generalizable evidence base.

CONCLUSIONS

In this scoping review, we have provided the first robust catalogue of LLM methodology in OHNS with a focus on reproducibility. Reporting of methodology in accessing and prompting LLMs is variable and frequently poor, leading to substantial questions regarding the generalizability of the vast majority of the reviewed studies. Careful attention by researchers and journals within OHNS towards higher standards in methodology reporting will be crucial to ensure reproducibility, strengthen the evidence base surrounding LLMs in OHNS and to eventually leverage these tools for patient benefit.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

John Usseglio, MPH, Senior Informationist at Columbia University Augustus C. Long Health Sciences Library, for assistance with search strategy creation.

FINANCIAL DISCLOSURES:

Anaïs Rameau was supported by a Paul B. Beeson Emerging Leaders Career Development Award in Aging (K76 AG079040) from the National Institute on Aging and by the Bridge2AI award (OT2 OD032720) from the NIH Common Fund.

Anaïs Rameau is an advisor for Sound Health Systems, Inc and for Pentax Medical.

REFERENCES

1. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023; 3:141. [PubMed: 37816837]
2. Zhang L, Zhao Q, Zhang D, Song M, Zhang Y, Wang X. Application of large language models in healthcare: A bibliometric analysis. *Digit Health* 2025; 11:20552076251324444. [PubMed: 40035041]
3. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025; 25:117. [PubMed: 40055694]
4. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 2025; 333:319–328. [PubMed: 39405325]
5. Lechien JR, Rameau A. Applications of ChatGPT in Otolaryngology-Head Neck Surgery: A State of the Art Review. *Otolaryngol Head Neck Surg* 2024; 171:667–677. [PubMed: 38716790]
6. Bany N, Ma B, Amanian A, Bur A, Abdalkhani A. Applications of Natural Language Processing in Otolaryngology: A Scoping Review. *Laryngoscope* 2025.
7. Zaghir J, Naguib M, Bjelogrić M, Neveol A, Tannier X, Lovis C. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *J Med Internet Res* 2024; 26:e60501. [PubMed: 39255030]
8. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Med Inform* 2024; 12:e55318. [PubMed: 38587879]
9. Wang J, Shi E, Yu S, et al. Prompt Engineering for Healthcare: Methodologies and Applications. Periodical [serial online]. Date 2023:arXiv:2304.14670. Available from. Accessed April 01, 2023.
10. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. Periodical [serial online]. Date 2024:arXiv:2402.07927. Available from. Accessed February 01, 2024.
11. Zhang X, Talukdar N, Vemulapalli S, et al. Comparison of Prompt Engineering and Fine-Tuning Strategies in Large Language Models in the Classification of Clinical Notes. *AMIA Jt Summits Transl Sci Proc* 2024; 2024:478–487. [PubMed: 38827053]
12. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit Med* 2024; 7:41. [PubMed: 38378899]
13. Meincke L, Mollick E, Mollick L, Shapiro D. Prompt Engineering is Complicated and Contingent Wharton AI & Analytics Initiative. Wharton AI & Analytics Initiative: Wharton School of the University of Pennsylvania, 2025.
14. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med* 2021; 13.

15. Nori H, Usuyama N, King N, et al. From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond. Periodical [serial online]. Date 2024;arXiv:2411.03590. Available from. Accessed November 01, 2024.
16. Reasoning best practices. Available at: <https://platform.openai.com/docs/guides/reasoning-best-practices>. Accessed June 7 2025.
17. Drog J, Milota M, van den Brink A, Jongsma K. Ethical guidance for reporting and evaluating claims of AI outperforming human doctors. NPJ Digit Med 2024; 7:271. [PubMed: 39358556]
18. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025; 31:60–69. [PubMed: 39779929]
19. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist: A Technical Report. NEJM AI 2025; 2.
20. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018; 169:467–473. [PubMed: 30178033]
21. Busch F, Hoffmann L, Rueger C, et al. Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. medRxiv 2024:2024.2003.2004.24303733.
22. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024; 11:1477898. [PubMed: 39534227]
23. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014; 96:395–403. [PubMed: 24973195]
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999; 53:105–111. [PubMed: 10396471]
25. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). Eur Arch Otorhinolaryngol 2023.
26. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? Periodical [serial online]. Date 2023;arXiv:2307.09009. Available from. Accessed July 01, 2023.
27. Hou Y, Bert C, Gomaa A, et al. Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. Front Artif Intell 2024; 7:1493716. [PubMed: 39877751]
28. Nazar W, Nazar G, Kaminska A, Danilowicz-Szymanowicz L. How to Design, Create, and Evaluate an Instruction-Tuning Dataset for Large Language Model Training in Health Care: Tutorial From a Clinical Perspective. J Med Internet Res 2025; 27:e70481. [PubMed: 40100270]
29. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023; 620:172–180. [PubMed: 37438534]
30. Maharjan J, Garikipati A, Singh NP, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Sci Rep 2024; 14:14156. [PubMed: 38898116]
31. He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. Does Prompt Formatting Have Any Impact on LLM Performance? Periodical [serial online]. Date 2024;arXiv:2411.10541. Available from. Accessed November 01, 2024.
32. Asgari E, Montana-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025; 8:274. [PubMed: 40360677]
33. Mesko B Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res 2023; 25:e50638. [PubMed: 37792434]
34. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol 2024; 108:1371–1378. [PubMed: 37923374]
35. Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ, Everson J. Current Use And Evaluation Of Artificial Intelligence And Predictive Models In US Hospitals. Health Affairs 2025; 44:90–98. [PubMed: 39761454]

36. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med* 2022; 5:194. [PubMed: 36572766]
37. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024; 30:1134–1142. [PubMed: 38413730]
38. Chen K, Stotter C, Klestil T, Nehrer S. Artificial Intelligence in Orthopedic Radiography Analysis: A Narrative Review. *Diagnostics (Basel)* 2022; 12.
39. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. *Nat Med* 2025.
40. Crowson MG, Rameau A. Standardizing Machine Learning Manuscript Reporting in Otolaryngology-Head & Neck Surgery. *Laryngoscope* 2022; 132:1698–1700. [PubMed: 35748581]

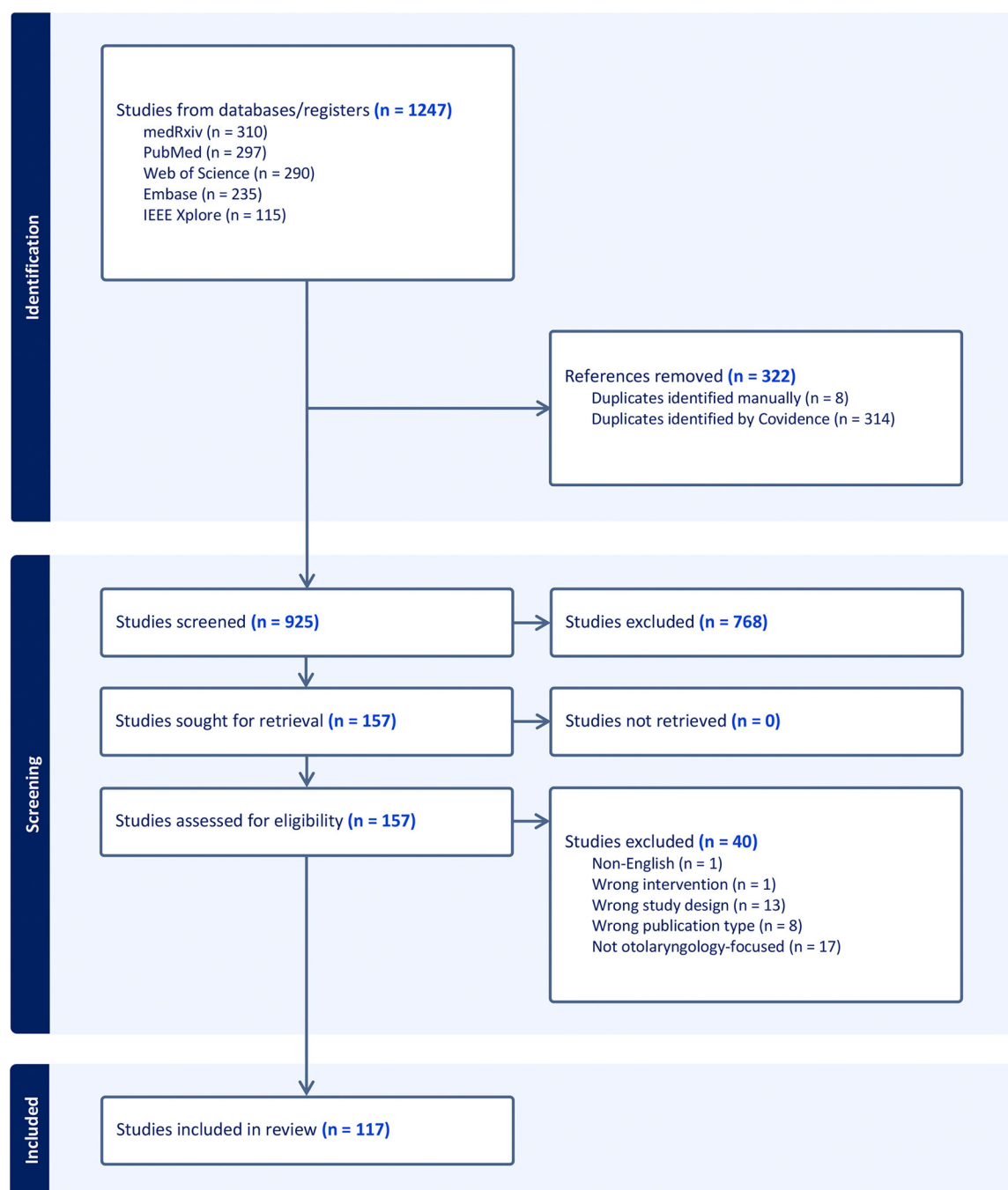
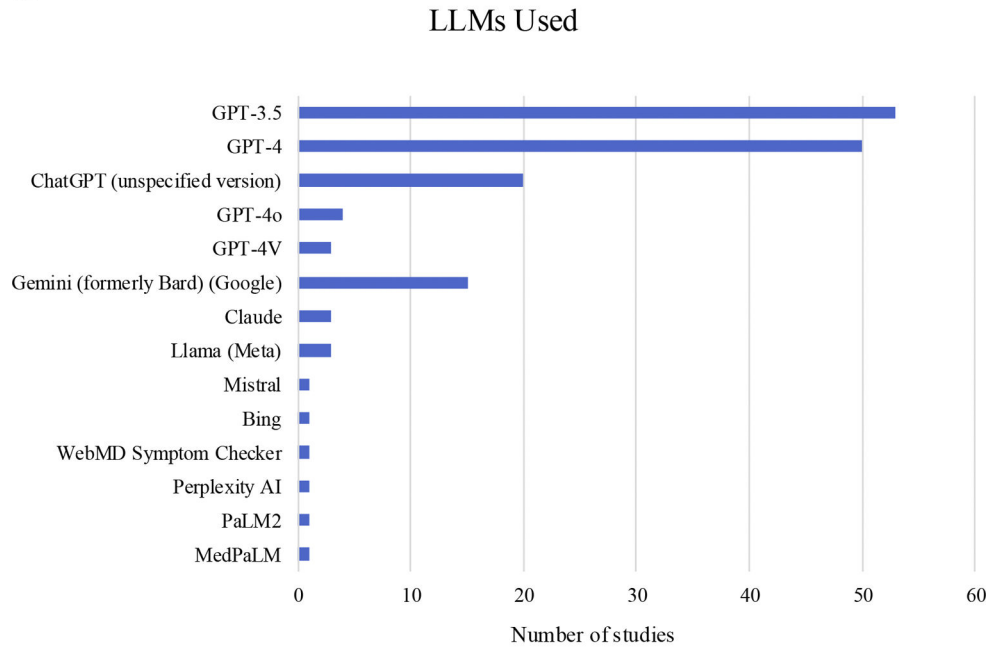
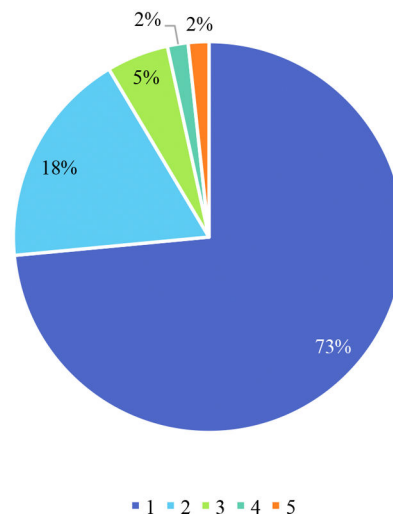
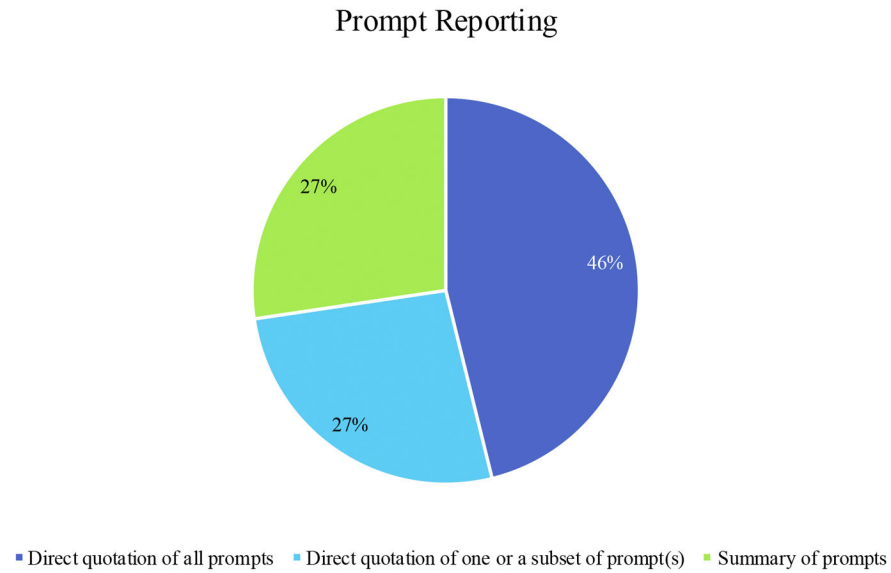
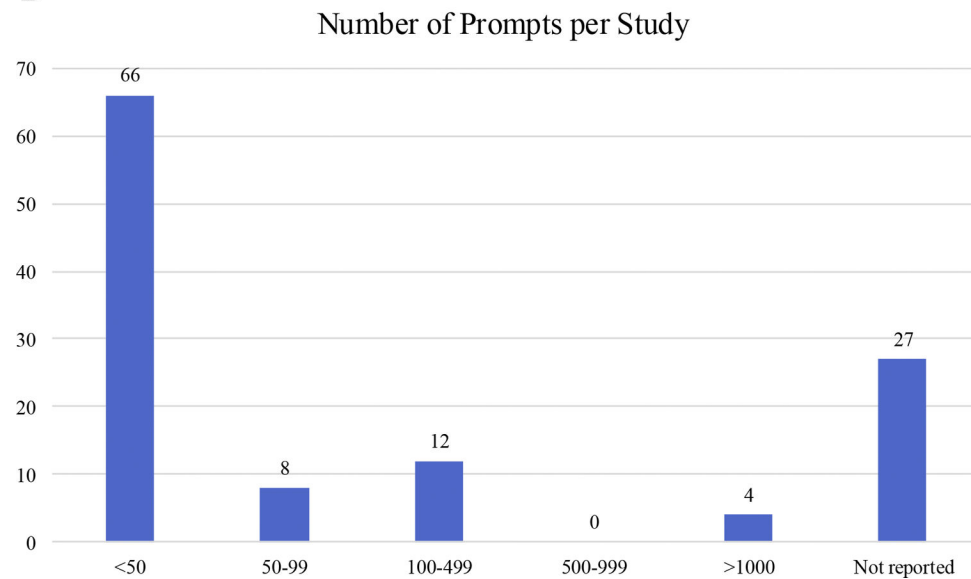


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISM-ScR) Flow Diagram.

A**B****Number of LLMs per Publication****Figure 2.**

Graphs showing a) number of included studies using each LLM and b) percentage of studies using or comparing a given number of LLMs.

A**B****Figure 3.**

Graphs showing a) degree of reporting of large language model prompts and b) number of studies using different numbers of prompts.