

Comparative evaluation of large language models for guideline-based oral cancer management: a standardized assessment of GPT-4o, DeepSeek-R1, and Grok-3

Jiayu Shen

Jilin University

Yao yuan

Jilin University

Sichen Han

Jilin University

Xinliang Duan

Jilin University

Peng Liu

Jilin University

Zilin Wang

wangzilin@jlu.edu.cn

Jilin University

Research Article

Keywords: oral cancer, oral cavity cancer, large language models, GPT-4o, DeepSeek-R1, Grok-3, NCCN guidelines, clinical decision support, artificial intelligence, oral oncology

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10904290/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) are increasingly being investigated as tools for clinical information synthesis, patient education, and decision support. However, fluent language generation does not necessarily translate into clinically accurate or guideline-concordant recommendations. The extent to which contemporary LLMs can translate evidence-based oral cancer guidance into patient-accessible information and individualized clinical management remains insufficiently characterized.

Methods

We developed a standardized evaluation framework based on the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology: Head and Neck Cancers, Version 2.2025. GPT-4o, DeepSeek-R1, and Grok-3 were evaluated in two tasks: guideline summarization for non-medical audiences and case-based oral cancer management. Model outputs were assessed by three oral and maxillofacial surgeons across predefined dimensions, including accuracy, clarity, conciseness, logical coherence, diagnostic accuracy, treatment-regimen appropriateness, surgical planning, radiotherapy, systemic-status assessment, and prognosis. BLEU-4 and ROUGE-L were treated as exploratory text-similarity metrics.

Results

In the guideline-summary task, DeepSeek-R1 achieved the highest reported overall assessment score (8.8/10), followed by GPT-4o (8.3/10) and Grok-3 (7.8/10). DeepSeek-R1 also achieved the highest reported scores for accuracy (9.0), clarity (8.7), conciseness (8.7), and logic (9.2). In the case-based task, DeepSeek-R1 achieved the highest reported comprehensive score (87.44), followed by Grok-3 (77.77) and GPT-4o (68.61). DeepSeek-R1 achieved high reported scores for treatment-regimen accuracy (92), surgical-resection accuracy (93), radiotherapy-dose assessment (98), and prognosis (97). Grok-3 achieved the highest reported radiotherapy-dose score (100). The raw dataset also contains several anomalous values that require verification before inferential analysis, including a reported ROUGE-L value of 207 and several clinical scores of 20.

Conclusions

The three LLMs demonstrated distinct performance profiles rather than a uniformly superior model. DeepSeek-R1 showed the strongest descriptive overall performance in the current benchmark, particularly for structured treatment planning and prognostic assessment. However, clinically relevant variability was observed in staging, lymph-node management, surgical extent, HPV contextualization,

and unsupported specificity. LLMs should therefore be considered assistive rather than autonomous clinical decision-support technologies. Larger multicentre studies using expert consensus standards and clinically meaningful endpoints are warranted.

Background

Cancer of the oral cavity is an important component of the global burden of head and neck malignancies and is associated with substantial morbidity, mortality, and functional impairment. Disease outcomes depend on early recognition, accurate pathological diagnosis, anatomical staging, appropriate treatment selection, and coordinated multidisciplinary care. Contemporary management may involve surgery, radiation therapy, systemic therapy, reconstruction, rehabilitation, nutritional management, and long-term surveillance according to tumour site, disease stage, and pathological risk factors.

The NCCN Clinical Practice Guidelines in Oncology: Head and Neck Cancers, Version 2.2025, provide evidence-based recommendations for cancers of the oral cavity and other head and neck sites. The guideline framework emphasizes site-specific and stage-specific treatment decisions and the integration of multidisciplinary care ^[1].

Translating complex clinical guidelines into accessible patient-facing information remains challenging because recommendations frequently depend on anatomical subsite, T and N classification, pathological findings, treatment intent, and risk factors. LLMs may help synthesize such information into structured natural-language outputs, but fluency alone does not establish clinical reliability.

GPT-4o is publicly described as an autoregressive multimodal model capable of processing text, vision, and audio ^[2]. DeepSeek-R1 is a reasoning-oriented model whose development emphasizes reinforcement learning for improved reasoning behavior ^[3]. Grok-3 has been described by xAI as a large-scale reasoning model trained using extensive computational resources and reinforcement learning ^[4]. These public descriptions characterize the models at a high level; they do not justify claims about undocumented proprietary internal modules.

In oral oncology, additional caution is required because oral cavity cancer and oropharyngeal cancer are anatomically and biologically distinct. HPV has a well-established role in oropharyngeal cancer, whereas it should not be indiscriminately treated as a universal driver of oral cavity cancer. Similarly, surgical extent, neck management, radiation treatment, and systemic therapy depend on the specific disease context ^[1].

Previous research has demonstrated potential applications of AI and LLMs in oncology and oral cancer diagnosis ^[5–7]. However, direct comparative evaluation of contemporary LLMs within a standardized oral oncology benchmark that integrates patient-facing communication and case-based clinical management remains limited. We therefore compared GPT-4o, DeepSeek-R1, and Grok-3 using a guideline-based two-

task framework and examined the relationship between expert clinical assessment and conventional text-overlap metrics.

Methods

1.Study design

This study was designed as a comparative benchmark evaluation of three LLMs for guideline-based oral cancer information synthesis and case-based clinical management. The evaluation consisted of a guideline-summarization task and a case-based management task. The workflow comprised guideline preparation, standardized prompting, response generation, independent expert evaluation, quantitative scoring, exploratory text-similarity analysis, and qualitative error analysis.

2.Guideline source

The primary clinical reference was the NCCN Clinical Practice Guidelines in Oncology: Head and Neck Cancers, Version 2.2025, with the analysis focused on oral cavity cancer and relevant head and neck treatment principles ^[1]. The exact guideline access date should be inserted before submission.

3.Models and experimental conditions

Three LLMs were evaluated: GPT-4o (OpenAI, USA), DeepSeek-R1 (DeepSeek, China), and Grok-3 (xAI, USA). The final submission should report the exact model identifier, access platform, testing date, browsing status, system instructions, generation parameters, and whether each model had equivalent access to the guideline. These details are essential for reproducibility because deployed LLMs may change over time.

4.Guideline-summarization task

All models received the same core prompt: “As a professional oral and maxillofacial surgeon and multidisciplinary expert, you are invited to review the following guideline. Your task is to summarize the oral cancer section from the Head and Neck Guidelines in a clear, concise, step-by-step explanation suitable for a non-medical audience. The summary should cover the following five areas: Etiology, Clinical Presentation, Diagnosis (including both the diagnostic basis and differential diagnosis), Treatment, Prognosis and Follow-Up.” The original model responses are retained in Supplementary Table S1.

5.Guideline-summary evaluation

Outputs were evaluated for accuracy, clarity, conciseness, logical coherence, and overall assessment on a 10-point scale. Accuracy addressed risk factors, diagnostic procedures, staging principles, treatment modalities, prognostic factors, and guideline concordance. Clarity assessed suitability for non-specialists and terminological precision. Conciseness assessed the balance between essential detail and unnecessary repetition. Logic assessed the clinical sequence from risk assessment to follow-up.

6. Case-based clinical management task

The case-based benchmark included standardized oral cancer scenarios representing mucosal lip, buccal mucosal, floor-of-mouth, tongue, and gingival carcinomas. Clinical materials included history, examination findings, imaging and pathological information where available, and relevant medical history. The exact number of cases must be reported after verification from the source dataset [TO BE INSERTED].

7. Clinical prompt

The models were instructed to provide a guideline-based clinical assessment including diagnosis, differential diagnosis, TNM staging where supported by available information, treatment modality selection, surgical procedure, extent of resection, lymph-node management, systemic-status assessment, radiotherapy considerations, postoperative care, and prognostic assessment.

8. Clinical evaluation domains

☒ Treatment-process regularity: coherence of diagnosis, staging, treatment, postoperative care, and follow-up.

☒ Diagnostic accuracy: correctness of anatomical site, disease classification, and staging information.

☒ Treatment-regimen accuracy: appropriateness of treatment selection under the supplied clinical context.

☒ Whole-body/systemic-status assessment: recognition of comorbidities, cardiopulmonary status, nutritional status, smoking-related risk, and treatment tolerance.

☒ Surgical-resection accuracy: appropriateness of the proposed surgical procedure, extent, and reconstruction.

☒ Radiotherapy assessment: appropriateness of indication, timing, target, and dose in context.

☒ Prognostic assessment: integration of stage, pathology, risk factors, recurrence risk, and follow-up.

9. Expert assessment

Three oral and maxillofacial surgeons independently assessed the model outputs. The final manuscript should state whether model identities were masked, whether output order was randomized, how disagreements were adjudicated, and whether reported scores represent means, medians, or consensus values. These details are not available in the current raw text and should be verified.

10. Statistical analysis

Descriptive statistics were used to summarize model performance. For repeated ordinal comparisons among the three models, the Friedman test was planned, followed by pairwise Wilcoxon signed-rank tests with correction for multiple comparisons if the global test was significant. The final manuscript should report the exact Friedman statistic, degrees of freedom, exact P values, adjusted pairwise P values, effect sizes, and confidence intervals where appropriate. The current aggregate tables do not contain enough information to derive these statistics reliably, so they are intentionally not fabricated here.

11. Text-overlap analysis

BLEU-4 and ROUGE-L were calculated as exploratory measures of similarity between generated and reference texts. These metrics were not considered direct indicators of clinical correctness. The final Supplementary Methods should define the precise implementation, preprocessing, reference-answer construction, and normalization procedures. The reported ROUGE-L value of 207 for GPT-4o in the radiotherapy-dose domain must be verified before submission.

Result

1. Guideline summarization

DeepSeek-R1 achieved the highest reported overall assessment score (8.8/10), followed by GPT-4o (8.3/10) and Grok-3 (7.8/10). DeepSeek-R1 also achieved the highest reported scores for accuracy (9.0/10), clarity (8.7/10), conciseness (8.7/10), and logic (9.2/10). These values are descriptive summaries of the available ratings; statistical significance should be reported only after analysis of the underlying rater-level data.

2. Qualitative guideline findings

GPT-4o generally produced clinically coherent summaries and correctly emphasized tobacco and alcohol exposure, clinical examination, biopsy, imaging, and major treatment modalities. However, the raw

outputs showed incomplete stage-specific treatment detail and required clearer distinction between oral cavity and oropharyngeal disease.

DeepSeek-R1 produced the most structured guideline summaries and more consistently incorporated multidisciplinary and stage-oriented concepts. Nevertheless, its outputs also included generalized statements about HPV, “genetic testing,” and radiotherapy de-escalation that require more precise disease-specific qualification.

Grok-3 generated concise and accessible responses with a logical sequence but showed greater limitations in disease-site specificity, HPV contextualization, and detailed stage-based treatment planning.

3. Case-based clinical management

The reported comprehensive clinical scores were 68.61 for GPT-4o, 87.44 for DeepSeek-R1, and 77.77 for Grok-3. DeepSeek-R1 achieved the highest reported comprehensive score. It also received high scores for treatment-regimen accuracy (92), surgical-resection accuracy (93), radiotherapy-dose assessment (98), and prognostic assessment (97). Grok-3 achieved the highest reported radiotherapy-dose score (100), while GPT-4o achieved a surgical-resection score of 91.

The raw dataset contains values of 20 for GPT-4o diagnostic accuracy and systemic-status assessment and 20 for Grok-3 systemic-status assessment. Because the surrounding metrics are reported on a 100-point scale, these entries require source-level verification before inferential analysis or final submission.

4. Case-specific observations

☒ Mucosal lip carcinoma

The models differed in staging and elective neck-management recommendations. Surgical recommendations generally converged on wide local excision with reconstruction, but margin descriptions ranged from 5–10 mm to ≥ 1 cm. These differences illustrate the need for case-specific expert adjudication.

☒ Buccal mucosal carcinoma

Responses ranged from extensive composite resection with multiple adjunctive procedures to more limited local resection. This heterogeneity suggests a risk of unsupported escalation of surgical extent when anatomical invasion is incompletely specified.

☒ Floor-of-mouth carcinoma

The models generally converged on wide local resection, neck management, reconstruction, and postoperative radiotherapy where indicated. Some responses, however, provided rigid chemotherapy or radiotherapy schedules without sufficiently explicit linkage to pathological risk factors.

☒ **Tongue carcinoma**

The models commonly recommended wide resection, neck management, reconstruction, and postoperative radiotherapy. Several outputs used highly specific resection extents without enough case-level evidence to support the exact surgical boundary.

☒ **Gingival carcinoma**

The models recognized advanced local disease and proposed maxillary resection and reconstruction. Some outputs additionally proposed detailed postoperative antibiotics, chlorhexidine, fluoride regimens, or HPV monitoring without clearly establishing that these recommendations were required by the case or reference guideline.

5.Exploratory text-similarity analysis

The reported METEOR composite score was 0.893 for DeepSeek-R1, 0.841 for GPT-4o, and 0.807 for Grok-3. In the domain-specific BLEU-4/ROUGE-L analysis, rankings differed across domains. For example, DeepSeek-R1 achieved the highest reported similarity for clinical diagnosis and choice of surgical procedure, whereas Grok-3 had the highest similarity for treatment-process regularity. This divergence indicates that textual similarity and clinical correctness are not interchangeable constructs.

Discussion

This study compared GPT-4o, DeepSeek-R1, and Grok-3 across two clinically relevant oral oncology tasks. The main finding was that performance varied across dimensions, and no model demonstrated universal superiority.

DeepSeek-R1 achieved the highest reported descriptive performance in both guideline summarization and case-based management. The observed strengths were concentrated in structured treatment planning, surgical planning, and prognosis. Importantly, these results describe observed task performance and should not be interpreted as evidence of a particular proprietary internal reasoning mechanism^[8].

GPT-4o displayed a more balanced profile, with relatively strong guideline summarization and high reported performance in surgical and radiotherapy domains. The raw data nevertheless contain anomalously low domain scores that must be verified. Until those values are confirmed, strong conclusions regarding GPT-4o's relative weakness in diagnosis or systemic assessment are premature.

Grok-3 performed well in treatment-process organization and radiotherapy-dose assessment. Its concise style may be useful for rapid preliminary communication. However, qualitative analysis identified limitations in anatomical contextualization and case-specific treatment tailoring.

A recurring issue across all three models was inappropriate transfer of information between oral cavity and oropharyngeal cancer. This is clinically relevant because HPV-associated disease, staging, prognosis, and treatment considerations are site specific^[1]. Future benchmark designs should explicitly test anatomical classification and penalize unsupported cross-site extrapolation.

The raw outputs also showed a second safety issue: unsupported specificity. Models sometimes generated exact surgical margins, neck dissections, chemotherapy schedules, antibiotic regimens, imaging intervals, rehabilitation measures, or HPV testing recommendations without sufficient case-specific evidence. Such outputs may appear authoritative while exceeding the information supported by the case or guideline.

The difference between expert assessment and text-overlap metrics has methodological importance. A clinically correct answer does not need to reproduce the wording of a reference answer, and a highly similar answer can still be clinically unsafe. Therefore, expert adjudication and guideline concordance should be primary endpoints in clinical LLM benchmarks, with BLEU-4 and ROUGE-L treated as secondary exploratory measures.

From a clinical implementation perspective, the findings support a human-in-the-loop framework[9–12]. LLMs may assist with guideline summarization, patient education, documentation, multidisciplinary preparation, and structured drafting. They should not independently determine staging, surgical extent, neck management, systemic therapy, or prognosis.

The study has several strengths, including a disease-specific benchmark, standardized core prompts, multiple contemporary models, multiple oral oncology scenarios, expert assessment, and raw-output qualitative analysis. Limitations include the limited case set, single-institution expert panel, lack of prospectively validated patient outcomes, sensitivity to model version and access conditions, and incomplete availability of rater-level statistics in the current dataset.

Before submission, the manuscript must undergo a final data audit. Specifically, the reported ROUGE-L value of 207 and the several domain scores of 20 should be reconciled with the raw computation sheets. The exact number of cases, rater-level scores, model identifiers, dates, browsing conditions, and scoring rubric must also be documented. These steps are essential for reproducibility and statistical validity.

Conclusions

GPT-4o, DeepSeek-R1, and Grok-3 demonstrated distinct performance profiles when translating oral cancer guidance into patient-facing summaries and case-based management recommendations. DeepSeek-R1 achieved the highest reported overall descriptive performance in the current dataset,

particularly in structured treatment planning and prognostic assessment. However, clinically relevant discrepancies persisted in staging, lymph-node management, surgical extent, anatomical context, and unsupported specificity. LLMs should therefore be regarded as assistive clinical technologies requiring expert verification rather than autonomous decision-makers. Larger multicentre benchmark studies based on expert consensus standards, transparent model configurations, and clinically meaningful endpoints are warranted.

Abbreviations

AI

artificial intelligence

BLEU

bilingual evaluation understudy

LLM

large language model

NCCN

National Comprehensive Cancer Network

NLP

natural language processing

ROUGE

recall-oriented understudy for gisting evaluation

TNM

tumour-node-metastasis.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Hospital of Stomatology, Jilin University (approval No. SJDKQ2026105). Only textual data from patient medical records were used; no identifiable images or photographs were included. All patient information was anonymized prior to analysis, precluding re-identification. Given the retrospective nature of the study and the absence of patient intervention or privacy infringement, the ethics committee granted a waiver of informed consent. The study was conducted in accordance with the Declaration of Helsinki and applicable ethical guidelines.

3.Availability of data and materials

The standardized prompts, scoring framework, anonymized model outputs, and aggregated evaluation data should be made available from the corresponding author on reasonable request, subject to institutional and data-governance requirements. The final supplementary files should include the

prompts, exact model identifiers, testing dates, clinical vignettes, scoring rubric, and de-identified rater-level data where permitted.

4.Competing interests

The authors declare that they have no competing interests.

Use of AI-assisted technologies during manuscript preparation

AI-assisted tools were used for language editing and manuscript preparation. All scientific content, data interpretation, references, and final text were independently reviewed and verified by the authors. The authors remain responsible for the accuracy and integrity of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 82301026) and the Graduate Innovation Fund of Jilin University (Grant No. 2026CX332, 2026CX338, 2026CX343).

Author Contribution

Jiayu Shen and Yao Yuan contributed to study conception, data acquisition, model evaluation, and manuscript preparation. Sichen Han and Duan Xinliang contributed to clinical interpretation and expert assessment. Peng Liu and Zilin Wang supervised the study and critically revised the manuscript. All authors should verify and approve the final CRediT taxonomy statement before submission.

Acknowledgements

Not applicable.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71:209–49.
2. Colevas AD, Cmelak AJ, Pfister DG, et al. NCCN Guidelines® Insights: Head and Neck Cancers, Version 2.2025. *J Natl Compr Canc Netw.* 2025;23:2–11.
3. Gallifant J, Fan L, Liao Z, Chen S, Aerts HJWL, Dligach D, Kouzy R, Bitterman DS. Evaluating the use of large language models to answer patient-facing clinical trial questions. *JCO Oncol Adv.*

2025;2:e2500137.

4. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;645:633–8.
5. Zhang Y, Yang D, Shi Y, Liu Y. Performance of large language models in lung cancer clinical decision-making: a comparative analysis based on DeepSeek, Grok, and GPT. *Cureus*. 2025;17(12):e99026.
6. Tobias MAS, Nogueira BP, Santana MCS, et al. Artificial intelligence for oral cancer diagnosis: What are the possibilities? *Oral Oncol*. 2022;134:106117.
7. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. *J Med Internet Res*. 2021;23:e29839.
8. Tang R, Zhang S, Ding C, et al. Artificial intelligence in intensive care medicine: bibliometric analysis. *J Med Internet Res*. 2022;24:e42185.
9. Nittas V, Ormond KE, Vayena E, Blasimme A. Realizing the promise of machine learning in precision oncology: expert perspectives on opportunities and challenges. *BMC Cancer*. 2025;25:276.
10. Wong EYT, Verlingue L, Aldea M, Franzoi MA, Umeton R, Halabi S, et al. ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP). *Ann Oncol*. 2025;36(12):1447–57.
11. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open*. 2023;6(11):e2343689.
12. Hao Y, Qiu Z, Holmes J, Löckenhoff CE, Liu W, Ghassemi M, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digit Med*. 2025;8:450.