

Metric-dependent evaluation of a retrieval-augmented large language model workflow for oral squamous cell carcinoma treatment planning: a retrospective study

Min Wu

Zhejiang University School of Medicine

Xiaozhe Li

Shandong University

Xindi Wei

Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University, National Clinical Research Center for Oral Diseases, Shanghai Research Institute of Stomatology

Huiyong Zhu

zhuhuiyong@zju.edu.cn

Zhejiang University School of Medicine

Research Article

Keywords: oral squamous cell carcinoma, large language model, retrieval-augmented generation, clinical decision support, treatment planning, clinical evaluation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10940900/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 evaluated for oncology decision support, but open-ended treatment planning poses a measurement problem: a clinically defensible recommendation may differ from a historical plan, and different automated metrics may reward different properties of the same output. Retrieval-augmented generation (RAG) may improve evidence grounding, but its contribution cannot be inferred without controlled ablation. We evaluated how four archived RAG-enabled LLM configurations agreed with clinician treatment plans for oral squamous cell carcinoma (OSCC) and whether apparent model ranking depended on the evaluation method.

Methods

This single-center retrospective evaluation used the same 100-case OSCC study set for four archived configurations (Claude, DeepSeek-R1, Doubao, and Qwen). Generated treatment plans were compared with archived clinician plans using a DeepSeek-based LLM-judge score, m3e-base embedding cosine similarity, ROUGE-1 recall, and an archived blinded five-physician rating protocol. Cohort descriptors were audited from the de-identified case register. Because complete case-level metric arrays and the case-by-rater matrix were not retained, model comparisons were summarized descriptively and no unsupported paired inference was reconstructed.

Results

The cohort had a mean age of 59.5 years (SD 13.6); 65% of cases were male and 69% were source-coded N0. Model ordering differed by automated metric: Qwen had the highest mean LLM-judge score (0.809; DeepSeek-R1, 0.807), Doubao had the highest mean embedding similarity (0.8489), and Claude had the highest mean ROUGE-1 recall (0.2729). No configuration was uniformly best. The archived clinician-rating summary showed a different descriptive pattern, but formal inter-rater reliability and paired model comparisons could not be recomputed from the available archive.

Conclusions

In this retrospective OSCC treatment-planning evaluation, apparent model performance was inseparable from the reference standard and evaluation method. The findings support feasibility of the combined RAG-enabled workflow and demonstrate the limitations of selecting a clinical LLM from a single aggregate metric; they do not establish clinical equivalence, autonomous safety, patient benefit, or a causal advantage of RAG. Future validation should use time-aligned evidence, locked system

configurations, independent clinical adjudication, paired case-level analyses, severity-weighted safety assessment, controlled RAG ablation, and clinician-in-the-loop prospective evaluation.

Background

Treatment planning for oral squamous cell carcinoma (OSCC) is a multidisciplinary decision problem rather than a single-label classification task. Management depends on primary-site anatomy, T and N category, resectability, surgical margins, extranodal extension, perineural and lymphovascular invasion, prior treatment, comorbidity, functional consequences, and patient preferences [1–5]. Even within guideline-concordant care, more than one defensible pathway may exist, particularly for management of the clinically negative neck and for postoperative radiotherapy or chemoradiotherapy. Consequently, an archived treatment plan is clinically meaningful as a record of care delivered at a particular time, but it is not necessarily a unique or error-free gold standard for evaluating an open-ended decision-support system.

Clinical LLM research is rapidly moving beyond examination-style benchmarks toward real patient cases, multidisciplinary decisions, and clinician-facing workflows. Recent oncology studies have shown that agreement with expert care varies by treatment domain, disease context, and the completeness of clinical information rather than remaining stable across tasks [16–20, 25–27]. In a 2026 breast-cancer study using 286 real multidisciplinary tumor-board cases, agreement was substantially lower for individualized surgical decisions than for several systemic-therapy decisions [26]. A subsequent real-world lung-cancer MDT evaluation showed that information completeness and input representation materially altered performance [27]. These observations are directly relevant to OSCC, where small differences in staging details, pathological risk factors, or treatment history can change the recommended sequence or intensity of therapy.

RAG provides one mechanism for constraining generation with external evidence [7]. Across biomedical studies, RAG has often improved average performance relative to unaugmented LLMs, but the magnitude and direction of benefit depend on the corpus, retrieval strategy, and evaluation design [8, 12, 14, 20–22]. A 2026 clinical RAG study further showed that structured corpora and hybrid retrieval can improve performance, whereas poorly organized retrieval may introduce context noise [29]. Evaluation creates a second source of uncertainty. Lexical-overlap, embedding-based, and LLM-as-a-judge approaches measure different properties of free-text outputs, and recent healthcare studies show that they can disagree with domain experts or with one another [10, 23, 24, 31, 32]. Thus, high semantic similarity does not guarantee that a safety-critical treatment decision is correct, while low lexical overlap may simply reflect an acceptable paraphrase or alternative formulation.

The wider field is therefore shifting from asking whether an LLM can generate a plausible answer to asking whether its output is reproducible, clinically acceptable, safe, useful to clinicians, and robust in real workflows. This progression is visible in safety-threshold studies [28], clinician-augmentation experiments [6, 33], and prospective shadow evaluations [34]. Against that background, the present

study evaluates four archived RAG-enabled LLM configurations on the same 100-case OSCC dataset. The primary objective is to characterize agreement with archived clinician treatment plans across complementary automated and physician-rating measures and to determine whether the apparent ordering of models changes with the evaluation lens. The study is deliberately framed as an evaluation of a combined, archived decision-support workflow; it does not attempt to prove a causal RAG effect, clinical equivalence, autonomous deployment readiness, or superiority of a single model family.

Methods

Study design, setting, and evaluation scope

This was a single-center retrospective clinical-technical evaluation of an existing 100-case OSCC study archive assembled at the Department of Stomatology, The First Affiliated Hospital, Zhejiang University School of Medicine, between 2020 and 2025. The unit of analysis was the clinical case. Four archived RAG-enabled LLM configurations were evaluated on the same case set, and case membership was not altered according to model performance. The present analysis audited the source cohort descriptors and interpreted the retained model-evaluation outputs without reconstructing unavailable case-level inference data.

The original study protocol required a recorded pathological diagnosis of OSCC, sufficient clinical and pathological information for treatment planning, and an archived clinician treatment plan. Study files used de-identified case identifiers rather than direct personal identifiers. The study was approved by the Ethics Committee of Zhejiang University School of Medicine (Approval No. 1396), which waived informed consent because of the retrospective design and use of de-identified data.

Study cohort and source-data audit

Age, sex, T category, N category, overall stage, recorded primary site, and histological differentiation were audited from the de-identified case register for descriptive characterization. Categories were reported as recorded in the source data; no retrospective restaging or forced anatomical reclassification was performed. Four records carried source-coded Tis/T0 or stage 0 entries and six carried a source-site label of tongue base. Because complete source pathology and staging-adjudication materials were not available for the present reanalysis, these records were retained as originally coded rather than reclassified after inspection of model results. Completeness of the descriptor fields used in Table 1 was assessed directly from the register.

Knowledge base construction and retrieval workflow

A curated knowledge base supported retrieval during treatment-plan generation. The source inventory in Additional file 1 contains 35 items: six guideline or guidance documents, one monograph, and 28

representative peer-reviewed papers relevant to head-and-neck or oral oncology. Retrieved material was used as contextual evidence for generation and was not treated as an autonomous clinical authority.

The documented retrieval pipeline included document normalization, segmentation, semantic retrieval, keyword-oriented retrieval, and result fusion before generation. Several replication-critical settings were not retained in the available archive, including immutable model or API build identifiers for every endpoint, exact chunk size and overlap, retrieval top-k, reranking configuration, complete prompt templates, random seeds, and the full set of generation parameters. These values were not reconstructed retrospectively. The clinical records span 2020–2025, whereas the knowledge-base inventory includes sources published through 2025; therefore, evidence available to the agent was not demonstrably time-matched to each historical clinician decision.

The source case register was Chinese-language, whereas the manuscript and displayed outputs are English. A complete case-level audit trail documenting translation, prompt language, retrieval language, and output language was not retained for every run. A translation-sensitivity analysis could therefore not be reproduced from the archive, and this limitation was considered when interpreting automated and physician-based evaluation.

LLM configurations and generation workflow

Model configurations

Four archived model labels were evaluated: Qwen-10B, Doubao-XL, DeepSeek-R1, and Claude-3. DeepSeek-R1 is a general-purpose reasoning model rather than an oncology-specific model [9]. Because immutable endpoint/build identifiers and complete model cards were not retained for all configurations, the comparison applies to the archived study configurations and should not be generalized to current or future versions of the corresponding model families.

Retrieval-augmented and structured generation

RAG and structured rationale prompting were used together. For each case, the retrieval stage surfaced context from the curated knowledge base and supplied it to the selected LLM, which generated a clinically organized rationale and treatment plan. No matched no-RAG or no-structured-rationale outputs were retained. The study therefore evaluates the combined workflow and cannot estimate the incremental contribution of retrieval or structured prompting as separate components.

The system accepted structured and narrative case information, retrieved related passages, and generated a treatment-planning response conditioned on that context. Explanatory text was treated as observable output rather than as a faithful representation of latent internal reasoning. Evidence

presentation was intended to improve traceability; clinical validity was determined by the evaluation procedures rather than by the presence of an explanation alone.

System architecture

The documented architecture used FastGPT, a web front end, a Python/FastAPI backend, MongoDB storage, Milvus vector search, PostgreSQL support, and an embedding component recorded as "text-embedding-3" (Fig. 1). Exact software versions and the complete deployment configuration were not retained. Figure 1 therefore documents the functional architecture of the archived system rather than a fully reproducible software specification.

The offline evaluation sequence comprised case input, evidence retrieval, response generation, and evidence presentation. Prototype web interfaces were used during development, but this retrospective study did not evaluate real-time deployment, workflow efficiency, regulatory compliance, treatment changes, or patient outcomes.

Reference standard and retrospective evaluation design

All four archived configurations were evaluated on the same 100 cases. For each case, the generated treatment plan was compared with the archived clinician plan. The clinician plan was prespecified as a pragmatic retrospective comparator rather than an error-free gold standard because OSCC management can permit more than one guideline-consistent pathway depending on pathological risk factors, neck-management strategy, prior therapy, comorbidity, functional trade-offs, and patient preference [3-5,13,18]. Accordingly, concordance is interpreted as agreement with documented care, not proof of absolute clinical correctness.

Expert clinical evaluation

The archived protocol specified blinded review by five attending physicians with more than 10 years of head-and-neck oncology experience. Reviewers were blinded to model identity and rated each recommendation on five 5-point Likert dimensions: diagnosis accuracy, guideline adherence, safety awareness, logical organization, and clinical utility. The five items were summed to a 5–25 composite score. Because all source records were already labeled as OSCC, the historical "diagnosis accuracy" item is reported as part of the archived rubric and is not interpreted as an independent diagnostic-validation endpoint. The most favorable Likert category is likewise treated as an ordinal rating rather than a validated deployment threshold. The complete case-by-rater matrix was not retained; therefore, clinician ratings are descriptive and formal intraclass correlation coefficient (ICC), kappa, rater-level confidence intervals, and test-retest reliability were not reconstructed.

Statistical analysis

Analyses were descriptive. Cohort characteristics were summarized as counts and percentages, with age additionally reported as mean, standard deviation (SD), median, and range. Agreement between generated and archived clinician plans was summarized using three retained automated measures: a DeepSeek-based LLM-judge score (0–1), m3e-base embedding cosine similarity, and ROUGE-1 recall [11]. These metrics capture different properties of generated text and were not treated as interchangeable clinical endpoints. The DeepSeek-based judge was not considered independent adjudication because a DeepSeek-R1 configuration was also among the candidate models, and contemporary evidence indicates that LLM-judge behavior can vary by model, criterion, language, and evaluator population [10,32]. Complete case-level metric arrays were not retained, precluding reproducible paired inferential analysis. Because all models had originally been evaluated on the same cases, any confirmatory reanalysis would need to preserve within-case pairing and report effect sizes and confidence intervals with prespecified control of multiple comparisons.

Results

Cohort characteristics

The evaluation set comprised 100 cases. Mean age was 59.5 years (SD 13.6), median age was 61 years (range 29–89), and 65 cases were male. Source-coded N0 disease was recorded in 69 cases. Overall stage was recorded as stage 0 in 4 cases, stage I in 23, stage II in 36, stage III in 21, stage IVA in 14, and stage IVB in 2. The most frequent recorded primary-site categories were oral tongue (43%), gingiva/alveolus (20%), and buccal mucosa (13%). No values were missing for the audited descriptor fields shown in Table 1. T category, N category, stage, and primary-site labels are reproduced as recorded in the source register without post-hoc restaging or anatomical reclassification.

Table 1. Characteristics of the 100-case evaluation set.

Characteristic	Value
Number of cases	100
Age, mean (SD), years	59.5 (13.6)
Age, median (range), years	61 (29-89)
Sex, male / female	65 / 35
T category: Tis / T0 / T1 / T2 / T3 / T4a	3 / 1 / 24 / 47 / 15 / 10
N category: N0 / N1 / N2b / N2c / N3b	69 / 25 / 3 / 1 / 2
Stage: 0 / I / II / III / IVA / IVB	4 / 23 / 36 / 21 / 14 / 2
Recorded primary site: oral tongue	43 (43%)
Recorded primary site: gingiva/alveolus	20 (20%)
Recorded primary site: buccal mucosa	13 (13%)
Recorded primary site: tongue base	6 (6%)
Recorded primary site: lip	6 (6%)
Recorded primary site: floor of mouth	4 (4%)
Recorded primary site: palate	2 (2%)
Recorded primary site: mixed labels	6 (6%)
Histological differentiation	Well-moderate 33; moderate 25; well 20; NOS SCC 9; moderate-poor 8; poor 3; other 2
Missingness in audited descriptor fields	0%

Note: Categories are reproduced from the audited source register without post-hoc restaging or forced anatomical reclassification.

Automated agreement metrics

The apparent ordering of models depended on the automated metric. Qwen had the highest mean DeepSeek-based judge score (0.809), narrowly above DeepSeek-R1 (0.807). Doubao had the highest mean m3e-base cosine similarity (0.8489), whereas Claude had the highest mean ROUGE-1 recall

(0.2729). Table 2 summarizes the retained aggregate results. Because the complete case-level score arrays were unavailable, no paired inferential comparison was reconstructed; the numerical differences are therefore descriptive and do not establish statistical or clinical superiority.

Table 2. Descriptive performance metrics across four archived model configurations.

Model	LLM-judge score, mean (SD)	m3e-base cosine similarity, mean (SD)	ROUGE-1 recall, mean (SD)
Claude	0.745 (0.147)	0.8459 (0.0239)	0.2729 (0.4067)
DeepSeek-R1	0.807 (0.151)	0.8482 (0.0263)	0.1445 (0.2430)
Doubao-XL	0.795 (0.163)	0.8489 (0.0255)	0.1404 (0.2215)
Qwen-10B	0.809 (0.126)	0.8487 (0.0264)	0.1444 (0.2543)

Note: The LLM-judge score was generated by the DeepSeek-based evaluator. Values are reproduced from archived result summaries; no new inferential testing was reconstructed from aggregate statistics.

Model-specific distributions

The model-specific distributions (Figs. 2–5) reinforce the aggregate finding that different evaluation lenses favor different configurations. Claude showed the highest lexical overlap with archived clinician plans but also wide dispersion in ROUGE-1 recall. DeepSeek-R1 had a high DeepSeek-based LLM-judge score, but that score is not independent because a DeepSeek-family model also served as the evaluator. Doubao-XL had the numerically highest mean m3e-base cosine similarity, whereas Qwen-10B had the numerically highest mean DeepSeek-based LLM-judge score. The separations were descriptive and cannot be interpreted as statistical or clinical superiority because the paired case-level arrays required for confirmatory comparison were not retained.

Multidimensional clinician evaluation

The archived clinician-review summary showed higher descriptive multidimensional ratings for DeepSeek-R1 and Claude than for Qwen and Doubao (Fig. 6), producing a different ordering from the three automated metrics. These ratings are exploratory because the complete case-by-rater matrix required for formal reliability analysis, paired model comparisons, and prespecified safety-threshold testing was not retained. No claim of statistical superiority or readiness for clinical use is therefore based on the clinician-rating summary.

The reviewer-level panel showed a visually similar ordering across the five physicians; however, visual parallelism is not evidence of inter-rater reliability. A formal assessment would require the original rating matrix and a prespecified ICC or kappa model with confidence intervals [15].

Discussion

The central finding is not that one archived LLM was best, but that the conclusion about which model appeared best changed with the evaluation method. Qwen ranked highest on the DeepSeek-based judge, Doubao on embedding similarity, and Claude on lexical overlap, while the physician-rating summary showed another descriptive pattern. In an open-ended treatment-planning task, this is a substantive clinical-informatics result rather than a minor measurement artifact. Recent real-world oncology studies likewise show that LLM performance varies across treatment domains, information states, and clinical contexts [25–27]. Our findings add a complementary point: even when the cases and generated outputs are fixed, model ordering can change because the evaluator and metric emphasize different properties of the response.

The reference standard is the first determinant of that interpretation. A historical clinician plan documents a decision made for a particular patient under the evidence, institutional practice, and preferences available at that time; it is not necessarily the only guideline-consistent option. Prior oncology work has therefore distinguished exact concordance from independently adjudicated acceptable alternatives [18, 25]. This distinction is especially important in OSCC, where neck management, postoperative radiotherapy or chemoradiotherapy, treatment sequencing, comorbidity, functional trade-offs, and patient preferences may support more than one defensible pathway [3–5, 13]. The temporal asymmetry in our archive adds another layer: the knowledge base includes sources through 2025, whereas the cases span 2020–2025. A recommendation for an earlier case may therefore be informed by evidence that post-dated the original clinician plan. Discordance should consequently trigger clinical adjudication and error characterization rather than being labeled automatically as model failure or clinician error.

The second determinant is the evaluation instrument itself. ROUGE-1 rewards lexical overlap and can penalize meaning-preserving paraphrases, whereas embedding similarity is more tolerant of wording differences but can remain high despite a clinically decisive change in management. Recent clinical-text studies and the S.C.O.R.E. framework have shown that lexical and semantic metrics do not consistently represent safety, context, reproducibility, or expert-perceived clinical quality [23, 24]. Large multicenter physician evaluations further demonstrate that human ratings vary with specialty, experience, and practice environment, and that AI evaluators cannot fully substitute for physician-centered assessment [32]. The DeepSeek-based judge in our study is additionally non-independent because a DeepSeek-family model was one of the candidates. Together, these findings argue against treating any single automated score as a deployment endpoint. For OSCC, evaluation should explicitly weight decision-critical content such as resectability, neck management, margin and extranodal-extension implications, indications for postoperative radiotherapy or chemoradiotherapy, contraindications, treatment sequencing, and potentially harmful recommendations.

RAG should be interpreted with the same discipline. Controlled biomedical and oncology studies support the potential of retrieval or context augmentation to improve evidence grounding [8, 20–22], but current

informatics research shows that benefit depends on corpus quality, retrieval strategy, and pipeline design [29]. Because the available archive contains only the combined RAG-plus-structured-generation workflow, the present results cannot identify whether retrieval, prompting, or the underlying model caused the observed agreement. The defensible conclusion is therefore feasibility of the archived workflow under its recorded configuration. A confirmatory experiment would need a versioned and time-aligned knowledge base, locked model builds and prompts, preserved retrieval traces, matched no-RAG and alternative-retrieval conditions, and case-level paired analysis before assigning a causal effect to RAG.

These methodological constraints also define the appropriate translational position of the system. The field is progressing from standalone model benchmarking toward evaluations of human-AI collaboration and real clinical workflow. A randomized trial showed that LLM assistance can improve physician performance on open-ended management tasks, although the effect was assessed in a simulated setting [6]. More recent work has evaluated clinician augmentation across experience levels [33] and used prospective shadow testing in emergency neurology [34]. In parallel, a 2026 BMC safety study evaluated predefined operating thresholds and found that performance sufficient for clinician-reviewed screening did not imply readiness for autonomous operation [28]. Our study sits earlier on this validation pathway: it evaluates archived model outputs against historical clinical decisions and expert ratings but does not test whether clinicians make better, safer, faster, or more guideline-complete decisions when using the system. The next clinically meaningful endpoint is therefore not higher text similarity; it is measurable improvement in clinician decision-making under supervised use.

Several limitations constrain the conclusions. The study was retrospective, single-center, and limited to 100 records without an external clinical cohort. The clinician plan was a pragmatic comparator rather than an adjudicated multi-expert gold standard, and retrospective summaries may not contain every pathological or preference-sensitive factor that influenced care. Four records carried source-coded Tis/T0 or stage 0 entries and six were labeled tongue base; they were retained because complete source adjudication materials were unavailable. The knowledge base was not demonstrably time-matched to each case. Several replication-critical model, retrieval, prompting, generation, and language-handling parameters were not retained. Complete case-level automated scores and the physician rating matrix were unavailable, preventing paired inference, independent-judge sensitivity analysis, and formal inter-rater reliability assessment. No controlled no-RAG or no-structured-generation ablation was available, and the evaluation was text-based rather than multimodal. Finally, the study did not assess prospective safety, severity-weighted clinical errors, workflow integration, treatment consequences, patient outcomes, or net clinical benefit.

Taken together, the study supports a staged validation framework that separates four questions: whether retrieval returns the right and time-appropriate evidence; whether generation produces a guideline-acceptable and safe plan; whether automated and human evaluators agree on clinically important errors; and whether use of the system improves clinician decisions or patient care. The present analysis addresses the second and third questions only indirectly through concordance and descriptive expert ratings. Its contribution is therefore methodological as much as technological: within

an archived RAG-enabled OSCC workflow, the reference standard, evaluator, and metric materially shape the apparent model conclusion. This finding aligns with the current direction of medical informatics toward multidimensional, clinically grounded, reproducible, and safety-aware evaluation before prospective deployment [23, 28, 30, 32–34].

Conclusions

In this single-center retrospective evaluation of 100 OSCC cases, four archived RAG-enabled LLM configurations showed metric-dependent agreement with clinician treatment plans, and no configuration was uniformly favored across automated and physician-rating measures. The divergence among LLM-judge, embedding-similarity, lexical-overlap, and exploratory clinician-rating signals demonstrates that model assessment in open-ended clinical decision support depends on both the reference standard and the evaluation method. The current evidence supports feasibility of the combined workflow but does not establish clinical equivalence, autonomous safety, patient benefit, or a causal advantage of RAG. Prospective multicenter validation should use time-aligned evidence retrieval, locked and reproducible system configurations, independent specialist adjudication, paired case-level statistics, formal rater-reliability assessment, severity-weighted safety analysis, controlled RAG ablation, and clinician-in-the-loop utility endpoints before clinical translation.

Abbreviations

API: application programming interface; ICC: intraclass correlation coefficient; LLM: large language model; MDT: multidisciplinary team; OSCC: oral squamous cell carcinoma; RAG: retrieval-augmented generation; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; SD: standard deviation.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Zhejiang University School of Medicine (Approval No. 1396). The requirement for informed consent was waived because of the retrospective design and use of de-identified study data.

Consent for publication

Not applicable.

Availability of data and materials

The de-identified data supporting the descriptive findings may be available from the corresponding author on reasonable request and subject to institutional and ethics requirements. Complete case-level model-output metric arrays and the original case-by-rater matrix were not retained in the available study archive and therefore cannot be provided for reproducible reanalysis.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

MW: Conceptualization, validation, data curation, formal analysis, writing—original draft, and writing—review and editing. XL: Methodology and writing—review and editing. XW: Validation. HZ: Writing—review and editing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Department of Stomatology, The First Affiliated Hospital, Zhejiang University School of Medicine, for technical and administrative support. The authors also acknowledge the patients whose de-identified clinical records contributed to this retrospective study.

References

1. Bugshan A, Farooq I. Oral squamous cell carcinoma: metastasis, potentially associated malignant disorders, etiology and recent advancements in diagnosis. *F1000Res*. 2020;9:229. 10.12688/f1000research.22941.1.
2. Kaur G, Yap T, Ramani R, McCullough M, Singh A. Assessing bias in the causal role of HPV in oral cancer: A systematic review and meta-analysis. *Oral Dis*. 2024;30(8):5379–87. 10.1111/odi.15062.
3. 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):2–11. 10.6004/jnccn.2025.0007.
4. Koyfman SA, Ismaila N, Crook D, et al. Management of the Neck in Squamous Cell Carcinoma of the Oral Cavity and Oropharynx: ASCO Clinical Practice Guideline. *J Clin Oncol*. 2019;37(20):1753–74. 10.1200/JCO.18.01921.

5. Machiels JP, Rene Leemans C, Golusinski W, Grau C, Licitra L, Gregoire V. Squamous cell carcinoma of the head and neck: EHNS-ESMO-ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2020;31(11):1462–75. 10.1016/j.annonc.2020.07.011.
6. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med.* 2025;31(4):1233–8. 10.1038/s41591-024-03456-y.
7. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv Neural Inf Process Syst.* 2020;33:9459–74.
8. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inf Assoc.* 2025;32(4):605–15. 10.1093/jamia/ocaf008.
9. DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv.* 2025;arXiv:250112948.
10. Williams G, Rutunda S, Nzabakira F, Mateen BA. Human evaluators vs. LLM-as-a-Judge: toward scalable evaluation of GenAI in global health. *npj Digit Med.* 2026. 10.1038/s41746-026-02992-w.
11. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out.* Association for Computational Linguistics; 2004. pp. 74–81.
12. Fink A, Rau A, Reiser M, Bamberg F, Russe MF. Retrieval-Augmented Generation with Large Language Models in Radiology: From Theory to Practice. *Radiol Artif Intell.* 2025;7(4):e240790. 10.1148/ryai.240790.
13. Jain PV, Mallick I, Manikantan K, et al. Prognostic triad: a novel method for decision-making for adjuvant treatment in stage I-II oral squamous cell carcinoma. *Int J Oral Maxillofac Surg.* 2025;54(1):23–30. 10.1016/j.ijom.2024.07.004.
14. Xu S, Yan Z, Dai C, Wu F. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Front Public Health.* 2025;13:1635381. 10.3389/fpubh.2025.1635381.
15. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016;15(2):155–63. 10.1016/j.jcm.2016.02.012.
16. Schmidl B, Hütten T, Pigorsch S, et al. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *Eur Arch Otorhinolaryngol.* 2024;281(11):6099–109. 10.1007/s00405-024-08828-1.
17. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. *Eur Arch Otorhinolaryngol.* 2025;282(9):4815–24. 10.1007/s00405-025-09524-4.
18. Ah-Thiane L, Heudel PE, Campone M, et al. Large Language Models as Decision-Making Tools in Oncology: Comparing Artificial Intelligence Suggestions and Expert Recommendations. *JCO Clin Cancer Inf.* 2025;9:e2400230. 10.1200/CCI-24-00230.

19. Prabhakaran S, Bell S, Lee JC, Kong JCH. The utility of large language models in oncological multidisciplinary team meetings: A systematic review. *Eur J Surg Oncol.* 2026;52(5):111741. 10.1016/j.ejso.2026.111741.
20. Hack S, Karni RJ, Maniaci A, et al. Evaluation of large language models as decision support tools for head and neck cancer management: A blinded multidisciplinary simulation study. *Oral Oncol.* 2026;174:107877. 10.1016/j.oraloncology.2026.107877.
21. Jun H, Tanaka Y, Johri S, et al. A context-augmented large language model for accurate precision oncology medicine recommendations. *Cancer Cell.* 2026;44(3):676–e6854. 10.1016/j.ccell.2025.12.017.
22. Vollmer A, Schorn L, Schrader F, et al. A Self-Controlled Benchmark of Retrieval-Augmented Generation for Large Language Models on Clinical Guideline Questions. *Diagnostics.* 2026;16(15):2456. 10.3390/diagnostics16152456.
23. Tan TF, Elangovan K, Ong J, et al. A S.C.O.R.E. framework for evaluating open-ended responses from large language models in healthcare. *Cell Rep Med.* 2026;7(7):102883. 10.1016/j.xcrm.2026.102883.
24. Dahlberg A, Käenniemi T, Winther-Jensen T, et al. Measuring the quality of AI-generated clinical notes: A systematic review and experimental benchmark of evaluation methods. *Artif Intell Med.* 2026;177:103421. 10.1016/j.artmed.2026.103421.
25. Yang K, Lee J, Jang JW, Sung PS, Han JW. Evaluating the clinical utility of large language models for hepatocellular carcinoma treatment recommendations: A nationwide retrospective registry study. *PLoS Med.* 2026;23(1):e1004855. 10.1371/journal.pmed.1004855.
26. Nazlı MA, Esmerer E, Keles A. Benchmarking large language models in breast cancer care: agreement with radiology-led multidisciplinary tumor board decisions. *BMC Med Inf Decis Mak.* 2026;26:243. 10.1186/s12911-026-03556-4.
27. Zhou Z, Wang M, Li J, et al. LLM-assisted multidisciplinary decision-making in lung cancer: efficacy of Gemini-2.5-Pro and GPT-5 in a real-world MDT setting. *BMC Med Inf Decis Mak.* 2026. 10.1186/s12911-026-03723-7.
28. Basu S, Patel SY, Sheth P, et al. AI safety evaluation in an underrepresented population: real-world performance of clinical decision support and frontier language models on Medicaid patient messaging triage. *BMC Med Inf Decis Mak.* 2026. 10.1186/s12911-026-03763-z.
29. Li P, Patel A, Vallamchetla SK, et al. Optimizing Retrieval-Augmented Generation (RAG) in clinical medicine: methods and performance evaluation. *J Am Med Inf Assoc.* 2026;33(8):1436–45. 10.1093/jamia/ocag056.
30. Bakken S. Improved evaluation frameworks are required to move application of LLMs from research into clinical practice. *J Am Med Inf Assoc.* 2026;33(3):551–2. 10.1093/jamia/ocag016.
31. Halici M, Salturk S, Sayin I, et al. A multidimensional benchmarking framework for large language models in oncologic decision making. *Sci Rep.* 2026;16:23086. 10.1038/s41598-026-61195-1.

32. Shi P, Li J, Yang Z, et al. Physicians and artificial intelligence diverge in evaluating large language models on real clinical cases. npj Digit Med. 2026. 10.1038/s41746-026-02942-6.

33. Tao Y, Lin X, Sun H, et al. Augmenting clinical decision-making with large language models: evaluation across general and specialty tasks. J Am Med Inf Assoc. 2026. 10.1093/jamia/ocag053.

34. Guo Y, Meng X, Yu E, et al. Development and prospective shadow evaluation of a domain-specific large language model for emergency neurological diagnosis. npj Digit Med. 2026;9:470. 10.1038/s41746-026-02644-z.

Figures

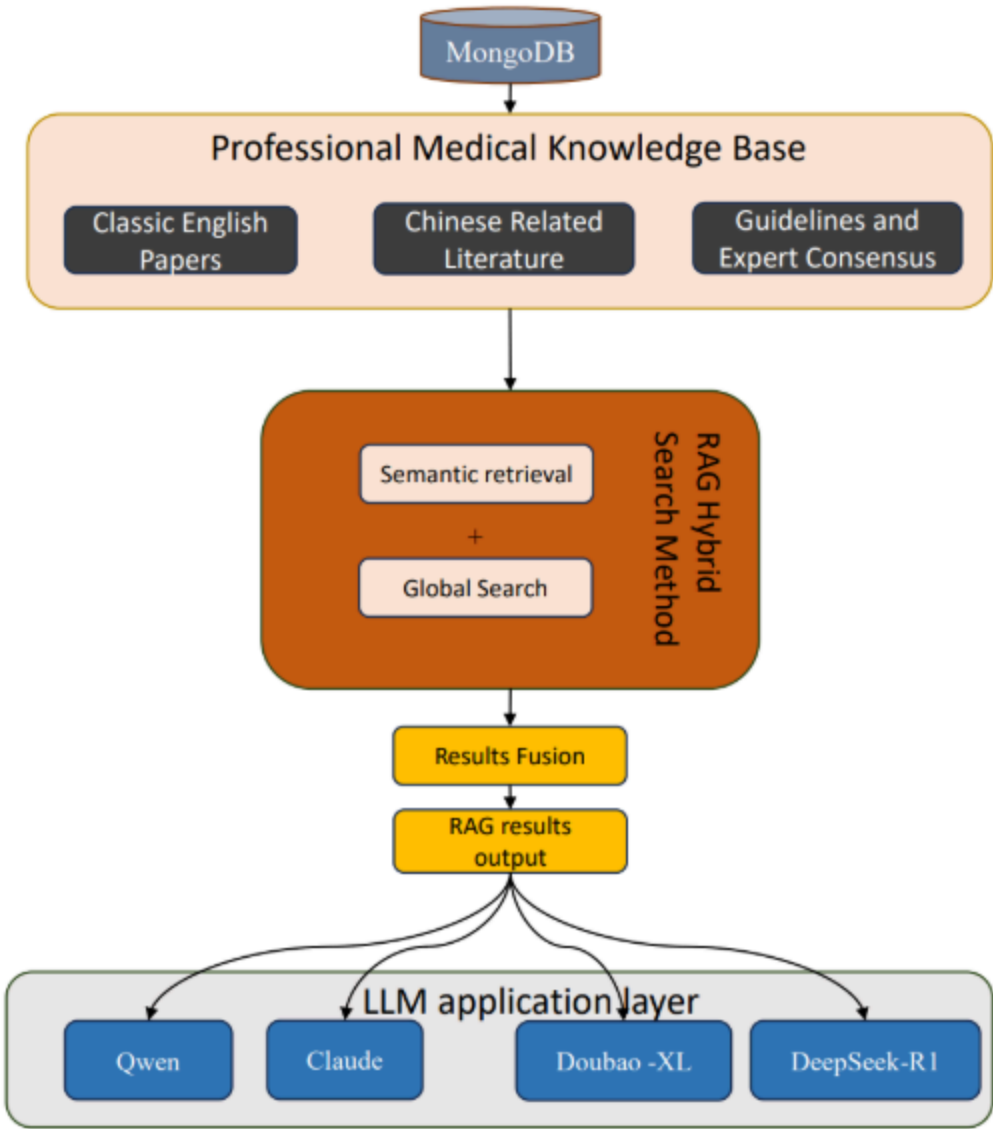


Figure 1

Functional architecture of the archived RAG-enabled OSCC treatment-planning system.

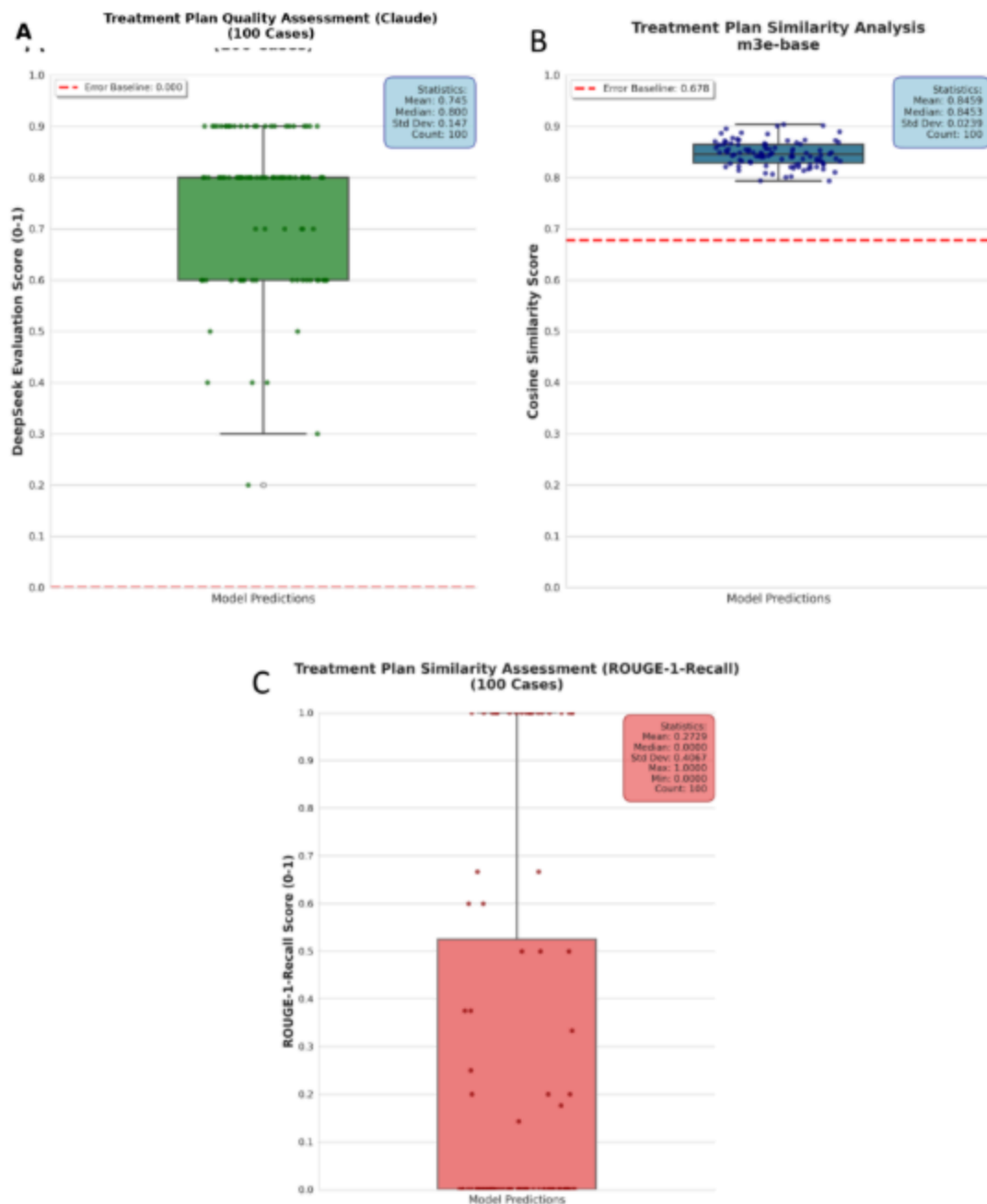


Figure 2

Archived automated evaluation results for the Claude configuration. (A) DeepSeek-based LLM-judge score. (B) m3e-base embedding cosine similarity. (C) ROUGE-1 recall. The automated judge is not an independent clinical adjudicator.

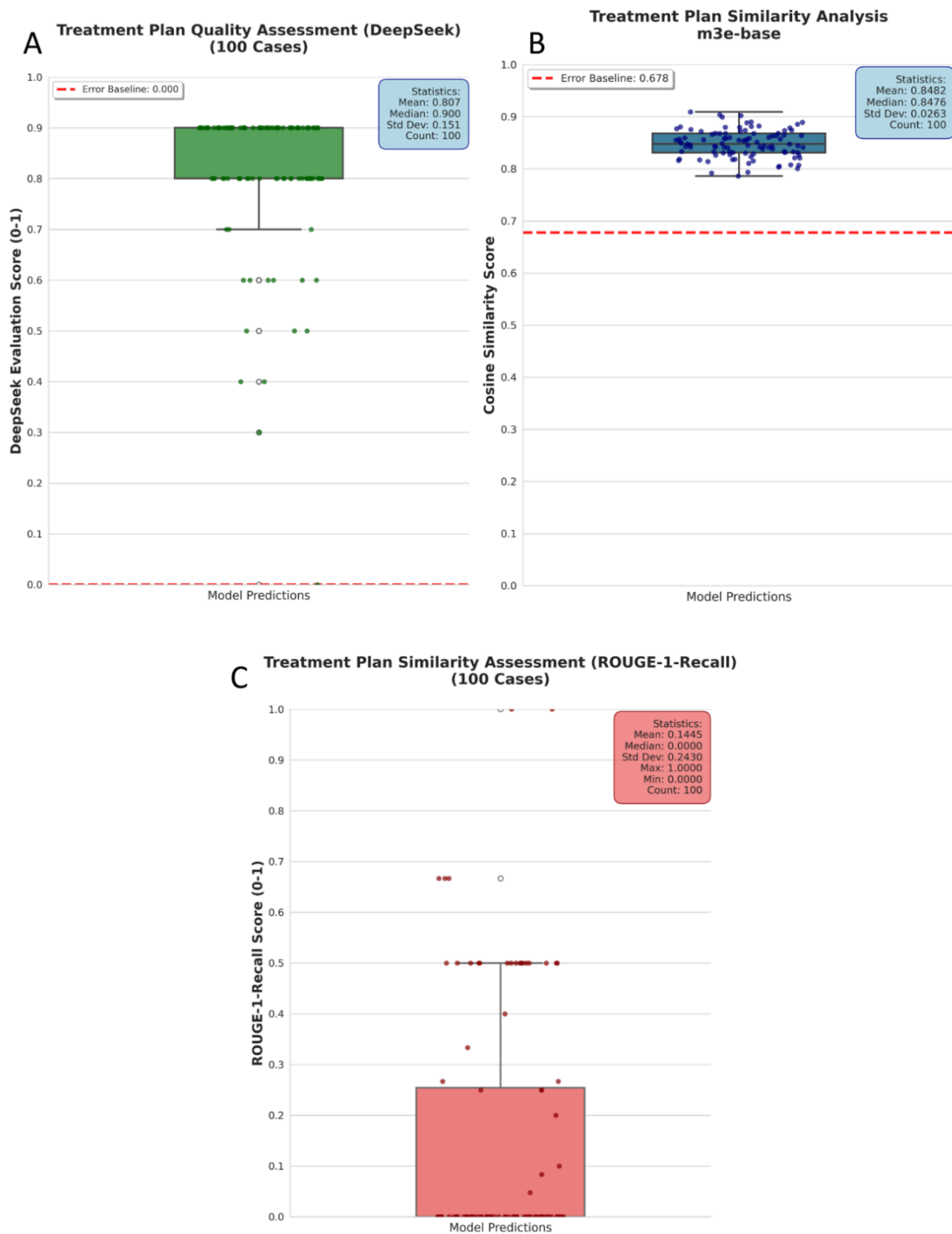


Figure 3

Archived automated evaluation results for the DeepSeek-R1 configuration. (A) DeepSeek-based LLM-judge score. (B) m3e-base embedding cosine similarity. (C) ROUGE-1 recall. The judge is potentially model-dependent because a DeepSeek-family model was also evaluated as a candidate.

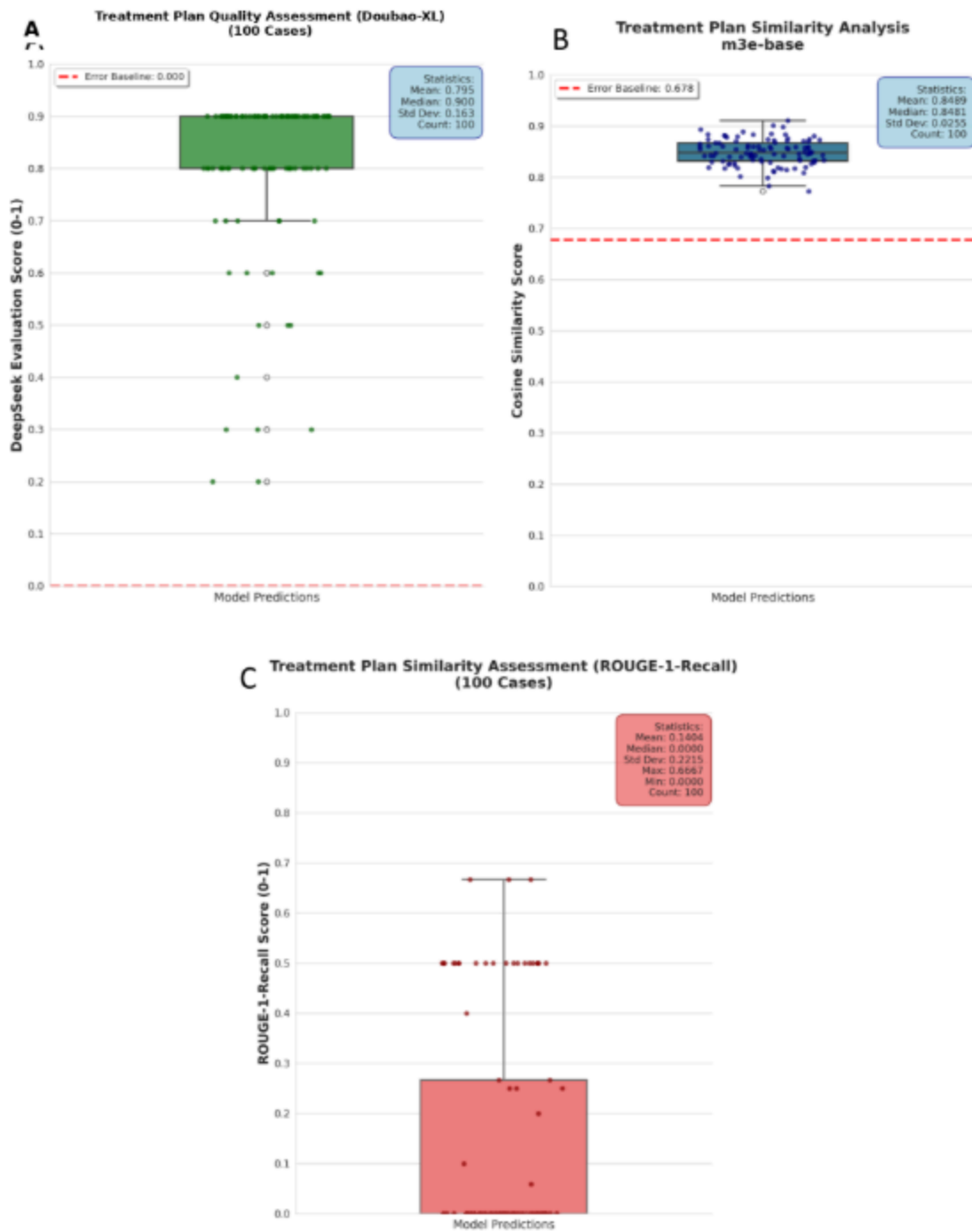


Figure 4

Archived automated evaluation results for the Doubao-XL configuration. (A) DeepSeek-based LLM-judge score. (B) m3e-base embedding cosine similarity. (C) ROUGE-1 recall.

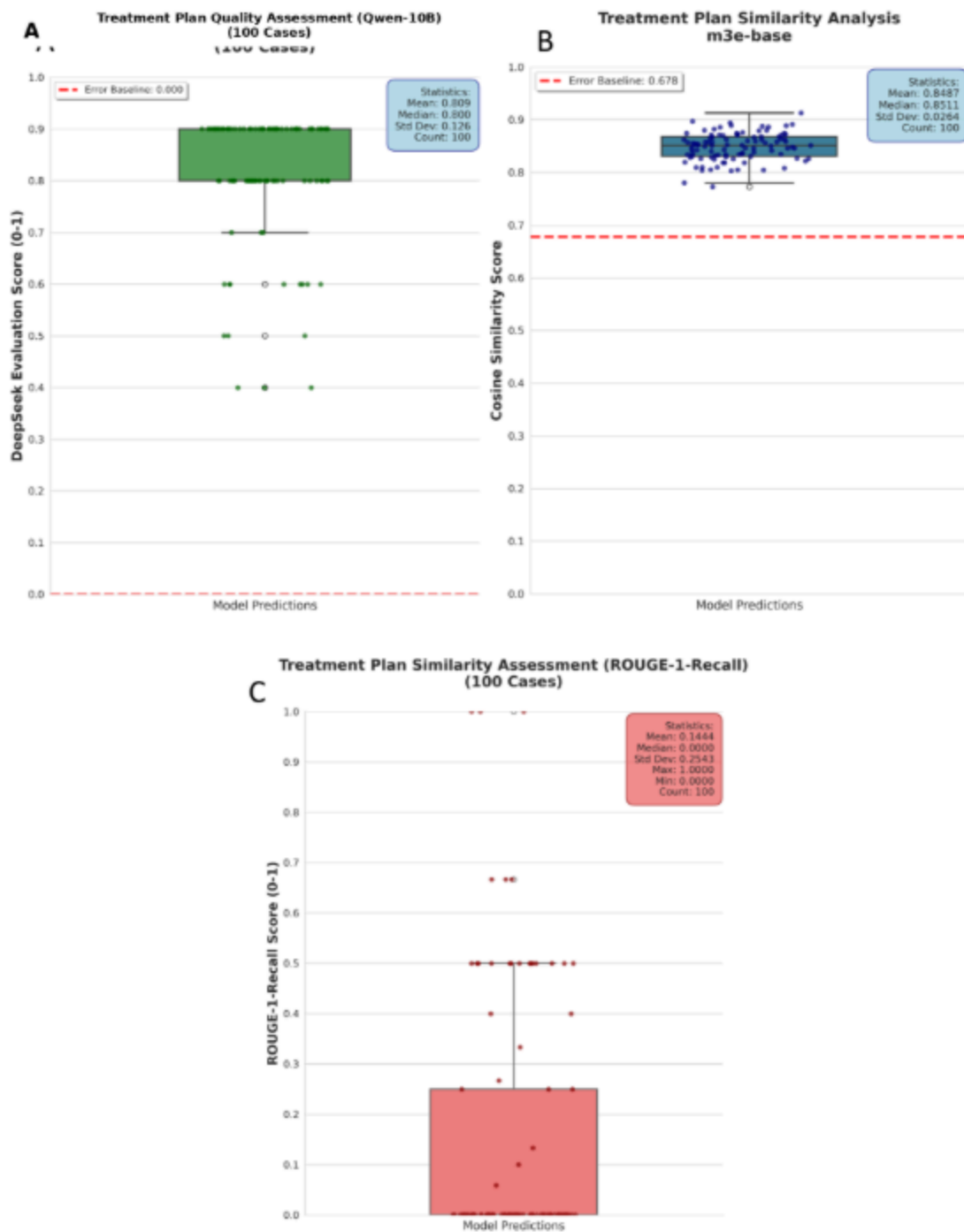


Figure 5

Archived automated evaluation results for the Qwen-10B configuration. (A) DeepSeek-based LLM-judge score. (B) m3e-base embedding cosine similarity. (C) ROUGE-1 recall.

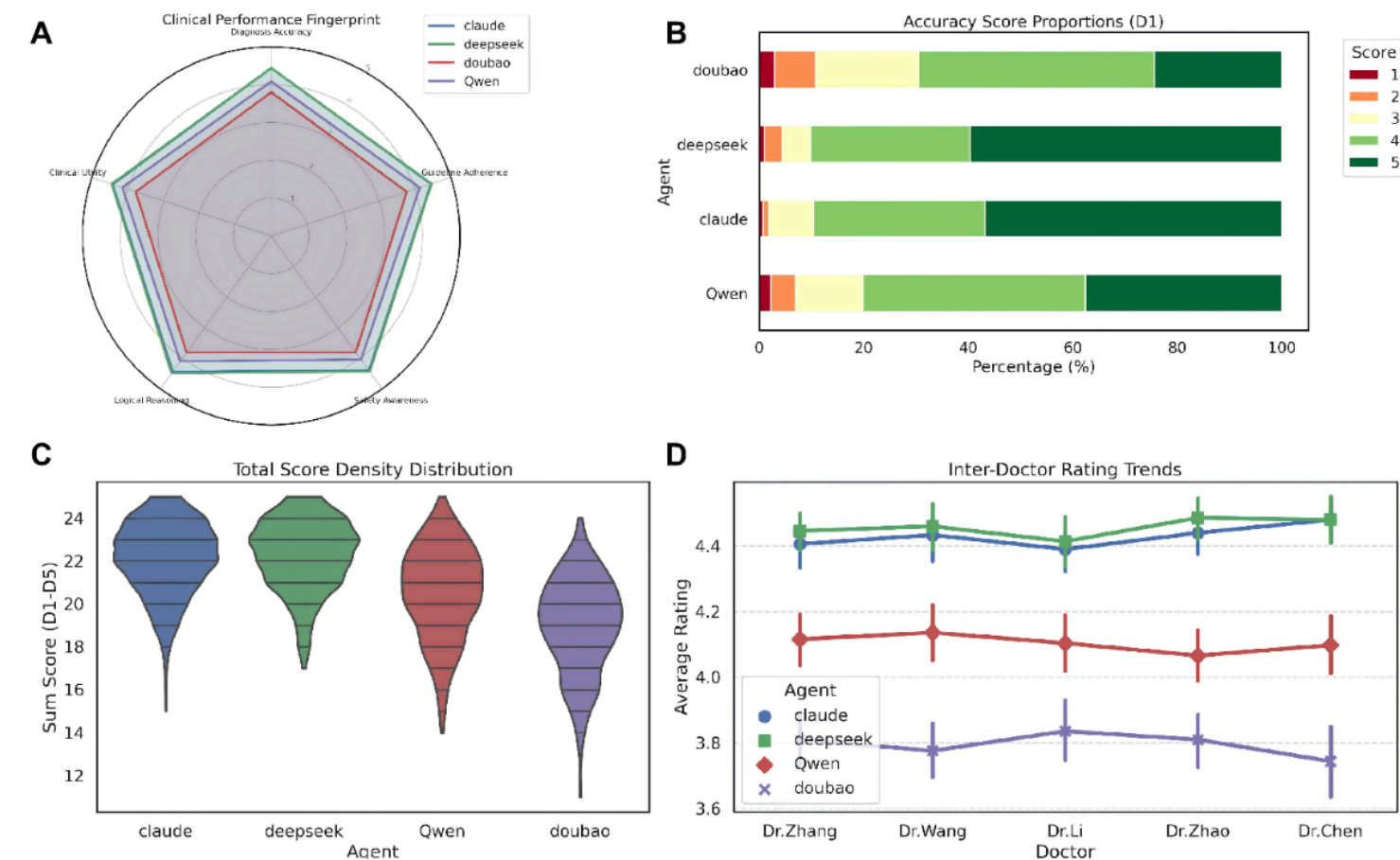


Figure 6

Exploratory clinician-rating summary for the four archived configurations. (A) Mean multidimensional rating profile. (B) Distribution of the historical diagnosis-accuracy Likert item. (C) Distribution of summed five-dimension scores. (D) Mean ratings by reviewer with error bars as displayed in the archived figure. Because the case-by-rater matrix was not retained, the figure is interpreted descriptively and does not establish inter-rater reliability or statistically significant model differences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable1.xlsx](#)