Evaluating Retrieval-Augmented Large Language Models on External Cervical Resorption: A Comparative Study of Gemini and NotebookLM

Abstract

Introduction: This study evaluated the accuracy and consistency of two large language models (LLMs) developed by Alphabet Inc., Google Gemini (GG), a base configuration, and NotebookLM (NLM), a document-grounded configuration, when answering clinical questions regarding external cervical resorption (ECR) using a retrieval-augmented framework. It was hypothesized that NLM would outperform GG due to its access to embedded reference documents.

Methods: Forty-six dichotomous clinical questions were developed by three academic endodontists based on established sources, including Heithersay's classification, Patel et al.'s reviews, and the European Society of Endodontology Position Statement. Each question was submitted to both models using three independent user accounts, yielding 276 responses. NLM was configured to generate responses exclusively from the provided documents, simulating a retrieval-augmented generation (RAG) setting. Three endodontic experts independently evaluated all responses against predefined gold standard answers. Accuracy was defined as agreement with the gold standard; consistency referred to identical responses across the three trials. Statistical analyses included 95% confidence intervals (Wald and Wilson), Fleiss' kappa, and Fisher's exact test.

Results: GG achieved an accuracy of 89% (41/46; 95% CI, 76.96–95.27) and a consistency rate of 93% ($\kappa = 0.89$; $p < 0.001$). NLM achieved an accuracy of 96% (44/46; 95% CI, 85.47–98.79) and the same consistency ($\kappa = 0.90$; $p < 0.001$). No significant differences were found (accuracy: $p > 0.05$; consistency: $p > 0.05$).

Conclusions: Both models demonstrated high accuracy and consistency. Although NLM showed slightly superior performance, retrieval augmentation did not significantly enhance outcomes in structured clinical tasks.

Keywords: external cervical resorption; large language models; retrieval-augmented generation; artificial intelligence; endodontics.

Introduction

External cervical resorption (ECR) is a pathological form of external tooth resorption characterized by clastic activity at the cervical region, typically beneath the junctional epithelium. This condition may progress asymptomatically and compromise surrounding dentin and pulp tissues (1–3). The unique histopathological features, etiological complexity, and frequent misdiagnosis of ECR distinguish it from other resorptive conditions (4,5). Early detection is essential but challenging, often requiring cone-beam computed tomography (CBCT) to visualize the lesion's full extent and guide appropriate intervention (6–9).

ECR lesions are commonly classified using Heithersay's four-grade system, which remains the most widely adopted clinical framework (10). Etiological factors include dental trauma, orthodontic treatment, periodontal surgery, and internal bleaching, with loss of cementum integrity emerging as a critical initiating factor (11–13). Treatment approaches vary depending on lesion severity and location and may involve surgical debridement, endodontic therapy, or extraction in advanced cases (14–17). The prognosis largely depends on the stage at diagnosis, the accuracy of lesion delineation, and the success of restoring defects (18,19).

To aid clinical decision-making in such complex scenarios, artificial intelligence (AI) tools, particularly large language models (LLMs), have emerged as promising allies. LLMs like Google Gemini (GG) and NotebookLM (NLM), both developed by Alphabet Inc., are capable of processing natural language input and generating contextually relevant clinical responses (20,21). Unlike traditional models, retrieval-augmented generation (RAG) architectures such as NLM combine generative reasoning with real-time document retrieval, producing responses grounded in specific user-provided references (22–24). This hybrid framework mirrors the Retrieval-augmented Information System for Enhancement (RISE)

structure and has demonstrated improvements in factual accuracy, reduced hallucination rates, and enhanced domain-specific performance in education and medical applications (24–26).

Despite the growing relevance of LLMs in clinical and academic settings, their application in endodontic conditions remains largely unexplored. To date, no prior study has examined whether retrieval-augmented models can outperform general-purpose LLMs in delivering guideline-concordant responses for rare or complex conditions such as ECR. This study introduces a novel comparative dimension by evaluating a retrieval-augmented model (NLM) alongside a baseline LLM (GG), thereby enabling the assessment of document-grounded reasoning capabilities in a dental context, an application not previously addressed in the LLM literature.

The present study aims to evaluate and compare the performance of two LLM configurations, GG (base) and NLM (RAG), in answering a series of expert-developed, guideline-based clinical questions related to ECR. By assessing accuracy and inter-account consistency across both models, this study seeks to determine whether retrieval augmentation improves the clinical reliability of AI-generated responses in ECR decision-making.

Materials and Methods

Ethics

This study involved simulated AI-generated content and expert evaluations and did not include human participants, interventions, or patient data.

Study Design and Scope

This was an observational, analytical, cross-sectional study conducted between May and July 2025. The aim was to evaluate and compare the accuracy and internal consistency of

responses generated by two configurations of LLMs developed by Alphabet Inc., GG (20) and NLM (21), when responding to evidence-based clinical questions related to ECR.

Question Development and Validation

A total of 46 dichotomous (Yes/No) questions were developed by three academic endodontists (F.D., F.A., M.G.-F.), each with over 10 years of clinical and academic experience in endodontics. These items were directly derived from the core clinical literature on ECR, including:

- Heithersay's classification system (10),
- Patel's diagnostic and therapeutic reviews (27, 28),
- The European Society of Endodontology Position Statement (29),
- A Q1-ranked clinical study on ECR management (30).

The questions were stratified across six domains**:** etiology, diagnosis, classification, treatment planning, prognostic factors, and material selection.

To ensure content validity, the 46 questions were piloted in a small internal group of endodontists (n = 3), who assessed each item for clarity, clinical relevance, and direct traceability to the reference material. Any ambiguous or compound questions were revised or eliminated. Each finalized question was linked to a single gold-standard answer, reviewed and agreed upon by consensus.

A full list of all questions and their gold-standard answers is available in **Supporting Information (Appendix 1)**.

Model Specifications and Configuration

Two LLMs developed by Alphabet Inc. were evaluated under distinct configurations:

- **GG** – *Base configuration (Flash 2.5 version)*

Accessed via the web interface (https://gemini.google.com/app) using a Google One subscription-based account**.** Each of the 46 questions was submitted through a new chat window, ensuring that no contextual memory or conversational history influenced subsequent answers. The model was queried without access to external documentation. No API, temperature adjustments, or system-level instructions beyond the platform's default settings were applied.

- **NLM** –*RAG configuration*

Accessed via https://notebooklm.google.com using a free Google account. The model was configured with embedded clinical reference documents in PDF format (references 10, 27–30), enabling document-grounded responses via the platform's native RAG architecture. This setup restricts outputs to the content provided in the uploaded sources.

Prompt design and execution were tailored to each model:

1. **GG:**
    - *Prompt:*
      "You are an experienced endodontist. Please answer the following clinical question about external cervical resorption (ECR) with either 'Yes' or 'No,' based only on current evidence-based guidelines and Q1-ranked journals. Provide a list ordered from 1 to 46 with no comments; only answer with 'Yes' or 'No.'"
2. **NLM:**
    - *Prompt:*
      "You are an experienced endodontist. Using the uploaded documents, answer the following clinical question about external cervical resorption (ECR) with either 'Yes' or 'No,' based only on the provided sources. Provide an ordered list from 1 to 46 with no comments. Only answer with 'Yes' or 'No.'"

No additional instructions or post-prompt interactions were permitted.

Each of the 46 clinical questions was submitted in a fixed order to both models, three independent times per model, using distinct user accounts (n = 276 responses in total). All submissions were conducted using fresh sessions (new chat or new document) to eliminate context carryover. This design simulates independent, real-world interactions with the AI systems. A schematic illustration of both model configurations is provided in **Figure 1**.

Consistency and Accuracy Assessment

Three independent academic endodontists, blinded to model identity, reviewed each of the 276 responses. Each response was evaluated as "accurate" or "inaccurate" by direct comparison with the predefined gold-standard answers.

Consistency was defined as the repetition of identical responses ("yes, yes, yes" or "no, no, no") across the three accounts per model for each question.

Final accuracy was computed based on the majority vote (2 of 3) per model per question. While this approach reduces granularity, it was chosen to simulate practical clinical consensus in low-repetition LLM testing.

The criteria and thresholds for accuracy and consistency, as well as the number of repetitions per question (n = 3), were selected to mirror the design of a prior validated study (31), ensuring methodological comparability while minimizing prompt fatigue and potential model adaptation effects.

A full list of all questions and their assessment is available in **Supporting Information (Appendix 2)**.

Statistical Analysis

Data were analyzed using R (version 4.3.1, 2023-06-16). Accuracy and consistency were reported as proportions, and 95% confidence intervals (CIs) were calculated using both the Wald and Wilson methods. Inter-account agreement was assessed using Fleiss's kappa coefficient. Comparative analyses between GG and NLM were performed using Fisher's exact

test, with statistical significance set at $p < 0.05$. This analytical framework enabled a structured comparison of performance between the two LLM configurations when addressing evidence-based clinical questions related to ECR.

To promote transparency and reproducibility, all raw data, evaluation sheets, and R scripts are available upon request.

Results

Google Gemini

A total of 46 dichotomous clinical questions were submitted to three independent GG accounts. Among the final consolidated responses, 34 (74%) were "YES" and 12 (26%) were "NO". Compared to the gold standard, comprising 31 "YES" (67%) and 15 "NO" (33%) expert-validated answers, GG achieved 41 accurate responses, yielding an accuracy of 89% (95% CI: 76.96%–95.27%; $p < 0.0001$, binomial test). Regarding internal agreement across accounts, 43 of 46 responses (93%) were consistent. Cohen's kappa coefficients ranged from 0.84 to 0.95 between account pairs, while Fleiss' kappa reached 0.89 ($z = 10.5$, $p < 0.001$), indicating strong inter-rater agreement.

NotebookLM

NLM responded to the same 46 questions using three independent accounts. The final outputs consisted of 32 "YES" (70%) and 14 "NO" (30%), closely matching the gold standard. The model produced 44 accurate responses, resulting in an accuracy of 96% (95% CI: 85.47%–98.80%; $p < 0.0001$, binomial test). Consistency was similarly high, with 43 of 46 answers (93%) being identical across accounts. Cohen's kappa values ranged from 0.84 to 0.95, and Fleiss' kappa reached 0.90 ($z = 10.5$, $p < 0.001$), again reflecting excellent inter-account reliability.

Item-Level Error Analysis

To gain deeper insight into the specific limitations of each LLM, an item-level error analysis was performed. Out of the 46 clinical questions, GG produced five incorrect answers, while NLM produced three. Notably, both models struggled with Question 17 ("¿In cases with a large buccal portal of entry in anterior teeth, is the external ECR approach the treatment of choice?"), incorrectly answering "Yes" despite the gold standard being "No", suggesting a shared vulnerability in interpreting nuanced surgical decision points.

GG displayed additional inconsistencies in questions related to radiographic interpretation (Q1), classification-based decisions (Q3, Q42), and historical outcome data (Q46). In contrast, NLM's incorrect responses mainly concerned clinical protocol prioritization and the interpretation of classification systems (Q17, Q25, Q42).

These findings highlight the importance of error analysis beyond aggregate metrics, revealing specific areas, such as classification comprehension and radiographic usage, where models may require further training or contextual grounding. **Table 1** summarizes the specific questions where at least one model deviated from the gold standard, including a breakdown of their responses and accuracy.

Comparative Analysis

Despite NLM achieving a higher point estimate for accuracy (96% vs. 89%) and slightly higher Fleiss' kappa (0.90 vs. 0.89), these differences were not statistically significant. Fisher's exact test showed no significant difference in accuracy between models (odds ratio = 0.38; 95% CI: 0.034–2.456; $p > 0.05$) or consistency (odds ratio = 1.0; 95% CI: 0.127–7.892; $p > 0.05$).

Thus, both LLM configurations demonstrated comparable performance in terms of internal consistency and alignment with expert-derived answers.

A summary of all statistical analyses is presented in **Table 2** and **Figures 2 and 3**.

Discussion

This study demonstrated that both Google Gemini (GG) and NotebookLM (NLM), two large language models (LLMs) developed by Alphabet Inc., delivered highly accurate and internally consistent responses to dichotomous clinical questions related to external cervical resorption (ECR). Although the retrieval-augmented model (NLM) achieved slightly higher accuracy (96%) than the base model (GG, 89%), the difference was not statistically significant. These results suggest that, under structured conditions and well-defined prompts, both configurations can serve as reliable adjunctive tools in endodontic education and evidence-based decision-making, aligning with prior applications of LLMs in medical and educational domains (22–26).

Retrieval-augmented generation (RAG) architectures, such as that of NLM, generate responses grounded exclusively in user-provided sources and have been associated with improved factual accuracy and reduced hallucination rates (24,26). However, our findings indicate that these advantages may be constrained when applied to narrowly framed, closed-ended clinical questions.

Our results align with prior studies reporting high NLM accuracy in clinical domains. Su et al. (32) compared ChatGPT-4o, a standard LLM, to an NLM supplemented with North American Spine Society guidelines, finding that NLM responses were significantly more accurate (98.3% vs. 40.7%, $p < 0.05$), more evidence-based (99.1% vs. 40.7%, $p < 0.05$), and more complete (94.1% vs. 79.7%, $p < 0.05$). The base model performance in the study of Su et al. (32) differed substantially from our GG results, which may reflect methodological differences, such as our use of dichotomous questions and models from the same parent company. Likewise, Cevik and Abu-Zidan (33) deployed a customized GPT model for thematic analysis in qualitative research and triangulated its outputs with NLM, observing strong alignment with manual coding. While human oversight remained necessary, their findings support the potential of retrieval-augmented frameworks to improve reliability in evidence-sensitive contexts.

The theoretical strength of RAG lies in its ability to tether outputs directly to verifiable sources, thereby improving factual integrity (22,24). Our study partially supports this: while both models performed well, NLM generated fewer factual errors, particularly in classification and diagnostic interpretation, echoing findings in oncology (26) and medical education (25), where document-grounded models have consistently outperformed base LLMs. However, the dichotomous nature of our task may have masked more nuanced advantages of retrieval augmentation, which are typically more evident in multi-step, interpretive tasks requiring contextual explanation or longitudinal reasoning.

Although the accuracy difference between NLM (96%) and GG (89%) did not reach statistical significance ($p > 0.05$), this 7% gap could still be clinically meaningful. In ECR, where diagnostic errors may result in unnecessary surgery or missed lesions (18,19), even small accuracy gains could influence outcomes. The lack of significance likely reflects the study's limited statistical power due to the modest sample size ($n = 46$). Future studies with larger datasets should examine whether such differences translate into consistent, clinically relevant advantages.

Item-level error analysis revealed shared limitations between GG and NLM on questions involving surgical decision-making (e.g., Q17), classification systems (Q42), and diagnostic imaging (Q1). These challenges likely stem from question complexity or clinical ambiguity rather than model architecture. These errors could be qualitatively categorized as arising from (a) misinterpretation of clinical nuance, such as overgeneralization of surgical thresholds; (b) ambiguity or under-specification in the source literature, particularly regarding radiographic staging or procedural guidance; and (c) confusion stemming from overlapping terminologies or classification systems. *This classification highlights targeted areas for future model refinement and prompt engineering.*

Notably, both models demonstrated strong internal agreement, with 93% consistency across accounts, reflecting stable generative behavior under controlled conditions and aligning with prior research on LLM reproducibility across users and sessions (31). Similar patterns have been observed in trauma-related evaluations, where base models like GG produced "consistently incorrect" answers, suggesting the presence of embedded, non-evidence-based

biases (31). In contrast, retrieval-augmented systems such as NLM generated fewer but more systematic errors, likely resulting from structured misinterpretation rather than hallucination, thus offering a safer margin for clinical oversight (31,34).

Methodologically, the study followed a controlled, reproducible evaluation framework consistent with validated approaches in LLM benchmarking (31,35). Nevertheless, certain limitations should be acknowledged. The binary question format, while useful for objective benchmarking, oversimplifies real-world clinical decision-making, which often involves conditional reasoning, uncertainty grading, and nuanced treatment considerations. Moreover, the limited number of repetitions per item (n = 3) restricts statistical power and may obscure variability due to temporal or contextual factors. Future research should incorporate more complex clinical scenarios, multi-turn dialogues, and longitudinal assessments to better capture model dynamics across time and user profiles. Beyond these methodological aspects, real-world ECR management requires sequential diagnostic and therapeutic reasoning. In actual clinical scenarios, decision-making often involves multiple interdependent steps, uncertainty grading, and context-specific reasoning. High performance in Yes/No questions does not imply competence in real-world tasks. Future studies should explore how LLMs perform in multifaceted or multi-turn tasks that better reflect the stepwise diagnostic and therapeutic challenges of managing ECR.

Additionally, both evaluated models originated from the same parent company, potentially limiting generalizability. This shared origin in developer and architecture may restrict the external validity of our findings, as performance patterns observed in GG and NLM might not extrapolate to models trained on different corpora, prompting strategies, or grounding mechanisms. Comparative studies involving LLMs from other developers (e.g., OpenAI, Anthropic, Mistral) and across multiple languages are needed to assess broader applicability and validate whether retrieval augmentation offers consistent advantages across platforms. Such research will be essential to determine whether the trends observed here, particularly regarding accuracy, consistency, and error analyses, are reproducible beyond the Alphabet ecosystem. Despite these limitations, this study contributes to the systematic evaluation of LLMs in endodontics, particularly in underexplored scenarios like ECR. Document-grounded

systems such as NLM may prove especially useful in settings with limited literature, elevated diagnostic uncertainty, and stringent guideline requirements.

It is also important to acknowledge that the endodontic literature, particularly concerning ECR, is more limited in both volume and scope compared with broader medical domains. This scarcity may have constrained the performance of NLM, whose outputs relied on a narrow pool of high-quality references. Although the selected documents represent the most authoritative sources currently available, the limited diversity of evidence may have attenuated the potential benefits of retrieval-based reasoning. Future research should explore whether similar trends are observed in more extensively documented conditions, such as apical periodontitis, where more heterogeneous guidelines exist.

In summary, this study adds to the growing body of evidence supporting the role of LLMs in specialized dental applications. While both GG and NLM performed admirably in terms of accuracy and consistency, the absence of statistically significant differences underscores the limitations of retrieval augmentation in narrowly scoped dichotomous tasks. Still, the slight accuracy advantage observed in NLM suggests that document-grounded reasoning could offer incremental benefits, especially in complex or low-prevalence clinical domains like ECR.

As LLMs become increasingly embedded in healthcare workflows, their adoption in dentistry should be guided by rigorous benchmarking, human-in-the-loop evaluation, and domain-specific validation. However, successful integration depends not only on technical performance but also on regulatory compliance and active professional engagement to ensure alignment with real-world clinical practice. As highlighted in trauma-related applications (31,34), AI-generated outputs must be treated as complementary tools, not as replacements for clinical expertise. Despite their strong quantitative performance, LLMs should not be relied upon as standalone diagnostic or decision-making tools. These models may generate inaccurate or misleading responses, particularly when faced with ambiguous or incomplete data inputs (36). Therefore, clinical use of LLMs must be accompanied by critical appraisal and human validation to prevent misinformation and ensure patient safety.

Future investigations should employ scenario-based vignettes, multi-turn prompts, and cross-linguistic comparisons to emulate real-world consultations and maximize clinical relevance. Such approaches will be critical to fully capturing the potential of retrieval-augmented architectures and clarifying their role in supporting diagnostic reasoning, patient education, and evidence-based decision-making in endodontic practice.

The practical implications of these findings are increasingly relevant as LLMs integrate into dental workflows. In endodontics, these tools may assist with clinical education, guideline consultation, and preliminary decision support, particularly in complex or low-prevalence scenarios like ECR. However, integration must include safeguards such as transparent sourcing, prompt standardization, and user training. Ethical concerns, including data privacy, bias propagation, and overreliance, must also be addressed. Crucially, LLMs should complement, not replace, clinical judgement. Their deployment should follow a human-in-the-loop framework, where outputs are critically appraised and verified by professionals before influencing patient care.

Beyond its empirical contributions, this study also provides a transparent, reproducible framework for AI evaluation in dentistry. By combining guideline-based question design, structured prompting, and multi-account replication, the methodology serves as a blueprint for benchmarking LLMs in other specialized domains, facilitating consistent evaluation as these technologies evolve.

Conclusions

Both GG and NLM exhibited strong performance when responding to structured, guideline-based clinical questions on ECR. While NLM´s retrieval-augmented configuration achieved slightly higher accuracy, the difference was not statistically significant, underscoring the limited added value of retrieval augmentation in closed-ended diagnostic tasks. Nonetheless, document-grounded LLMs may prove more beneficial in complex, context-rich scenarios that demand nuanced clinical reasoning, highlighting the need for continued investigation in endodontic research and practice.

Acknowledgments

The authors declare that they have no conflicts of interest related to this study. No financial affiliations, consultancy roles, stock holdings, or honoraria have influenced the conduct or reporting of this research. This study received no external funding, grants, contracts, or material support from any commercial or academic entity. All authors affirm that they have no financial involvement with any organization with a direct interest in the subject matter discussed in this manuscript, nor have any such arrangements existed in the past three years.

References

1. Heithersay GS. Clinical, radiologic, and histopathologic features of invasive cervical resorption. *Quintessence Int*. 1999;30(1):27–37. PMID: 10323156.

2. Ne RF, Witherspoon DE, Gutmann JL. Tooth resorption. *Quintessence Int*. 1999;30(1):9–25. PMID: 10323155.

3. Patel S, Foschi F, Mannocci F, Patel K. External cervical resorption: a three-dimensional classification. Int Endod J. 2018 Feb;51(2):206-214. doi: 10.1111/iej.12824.

4. Andreasen, J. O., F. M. Andreasen, and L. Andersson, eds. Textbook and Color Atlas of Traumatic Injuries to the Teeth. Fifth edition. Hoboken, NJ: Wiley-Blackwell, 2019.

5. Patel S, Durack C, Abella F, Shemesh H, Roig M, Lemberg K. Cone beam computed tomography in Endodontics - a review. Int Endod J. 2015 Jan;48(1):3-15. doi: 10.1111/iej.12270.

6. Durack C, Patel S. Cone beam computed tomography in endodontics. Braz Dent J. 2012;23(3):179-91. doi: 10.1590/s0103-64402012000300001.

7. Pereira SA, Corte-Real A, Melo A, Magalhães L, Lavado N, Santos JM. Diagnostic Accuracy of Cone Beam Computed Tomography and Periapical Radiography for Detecting Apical Root Resorption in Retention Phase of Orthodontic Patients: A Cross-Sectional Study. J Clin Med. 2024 Feb 22;13(5):1248. doi: 10.3390/jcm13051248.

8. Kaur K, Saini RS, Vaddamanu SK, Bavabeedu SS, Gurumurthy V, Sainudeen S, Mathew VB, Khateeb SU, Mokhlesi A, Mosaddad SA, Heboyan A. Exploring Technological Progress in Three-Dimensional Imaging for Root Canal Treatments: A Systematic Review. Int Dent J. 2025 Apr;75(2):1097-1112. doi: 10.1016/j.identj.2024.05.014.

9. Patel S, Brown J, Semper M, Abella F, Mannocci F. European Society of Endodontology position statement: Use of cone beam computed tomography in Endodontics. *Int Endod J.* 2019 Dec;52(12):1675–1678. doi:10.1111/iej.13187.

10. Heithersay GS. Invasive cervical resorption: an analysis of potential predisposing factors. Quintessence Int. 1999 Feb;30(2):83-95. PMID: 10356560.

11. Tronstad L. Root resorption--etiology, terminology and clinical manifestations. Endod Dent Traumatol. 1988 Dec;4(6):241-52. doi: 10.1111/j.1600-9657.1988.tb00642.x.

12. Galler KM, Grätz EM, Widbiller M, Buchalla W, Knüttel H. Pathophysiological mechanisms of root resorption after dental trauma: a systematic scoping review. BMC Oral Health. 2021 Mar 26;21(1):163. doi: 10.1186/s12903-021-01510-6.

13. Mavridou AM, Bergmans L, Barendregt D, Lambrechts P. Descriptive Analysis of Factors Associated with External Cervical Resorption. J Endod. 2017 Oct;43(10):1602-1610. doi: 10.1016/j.joen.2017.05.026.

14. Patel S, Kanagasingam S, Pitt Ford T. External cervical resorption: a review. J Endod. 2009 May;35(5):616-25. doi: 10.1016/j.joen.2009.01.015.

15. Talabani RM, Garib BT, Masaeli R, Zandsalimi K, Ketabat F. Biomineralization of three calcium silicate-based cements after implantation in rat subcutaneous tissue. Restor Dent Endod. 2020 Dec 2;46(1):e1. doi: 10.5395/rde.2021.46.e1.

16. Duarte MAH, Marciano MA, Vivan RR, Tanomaru Filho M, Tanomaru JMG, Camilleri J. Tricalcium silicate-based cements: properties and modifications. Braz Oral Res. 2018 Oct 18;32(suppl 1):e70. doi: 10.1590/1807-3107bor-2018.vol32.0070.

17. Talpos-Niculescu RM, Nica LM, Popa M, Talpos-Niculescu S, Rusu LC. External cervical resorption: Radiological diagnosis and literature (Review). Exp Ther Med. 2021 Oct;22(4):1065. doi: 10.3892/etm.2021.10499.

18. Irinakis E, Haapasalo M, Shen Y, Aleksejuniene J. External cervical resorption - Treatment outcomes and determinants: A retrospective cohort study with up to 10 years of follow-up. Int Endod J. 2022 May;55(5):441-452. doi: 10.1111/iej.13717.

19. Jeng PY, Chang SH, Wang CY, Lin LD, Jeng JH, Tsai YL. Surgical repair of external cervical resorption - Prognosis and prognostic factors. J Dent Sci. 2024 Jan;19(1):377-386. doi: 10.1016/j.jds.2023.08.005.

20. Google. Gemini (Internet). Mountain View (CA): Google; (cited 2025 Apr 30). Available from: https://gemini.google.com/app

21. Google. NotebookLM: note-taking & research assistant. Powered by AI (Internet). Mountain View (CA): Google; (cited 2025 Apr 30). Available from: https://notebooklm.google/

22. Yuan M, Bao P, Yuan J, et al. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Med Plus*. 2024;2(2):100030. doi:10.1016/j.medp.2024.100030

23. Yang X, Li T, Su Q, et al. Application of large language models in disease diagnosis and treatment. *Chin Med J*. 2024;138(1):e240001. doi:10.1097/CM9.0000000000003456

24. Yeo M, Moorhouse B, Wan Y. From academic text to talk-show: deepening engagement and understanding with Google NotebookLM. *TESL-EJ*. 2024;28(1):ej112int. doi:10.55593/ej.28112int

25. Mehta N, Agrawal A, Benjamin J, et al. Pedagogy and generative artificial intelligence: applying the PICRAT model to Google NotebookLM. *Med Teach*. 2024;46(6):721–8. doi:10.1080/0142159X.2024.2418937

26. Tozuka R, Johno H, Amakawa A, et al. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. *Jpn J Radiol*. 2024;42:901–7. doi:10.1007/s11604-024-01705-1

27. Patel S, Mavridou AM, Lambrechts P, Saberi N. External cervical resorption-part 1: histopathology, distribution and presentation. Int Endod J. 2018 Nov;51(11):1205-1223. doi: 10.1111/iej.12942.
28. Patel S, Foschi F, Condon R, Pimentel T, Bhuva B. External cervical resorption: part 2 - management. Int Endod J. 2018 Nov;51(11):1224-1238. doi: 10.1111/iej.12946.
29. European Society of Endodontology (ESE); Patel S, Lambrechts P, Shemesh H, Mavridou A. European Society of Endodontology position statement: External Cervical Resorption. Int Endod J. 2018 Dec;51(12):1323-1326. doi: 10.1111/iej.13008.
30. Espona J, Roig E, Durán-Sindreu F, Abella F, Machado M, Roig M. Invasive Cervical Resorption: Clinical Management in the Anterior Zone. J Endod. 2018 Nov;44(11):1749-1754. doi: 10.1016/j.joen.2018.07.020.
31. Dufey Portilla N, Garcia-Font M, Nagendrababu V, Abbott PV, Gonzalez Sanchez JA, Abella F. Accuracy and consistency of Gemini responses regarding the management of traumatized permanent teeth. Dent Traumatol. 2025 Apr;41(2):171-177. doi: 10.1111/edt.13004.
32. Su AY, Knebel A, Xu AY, Kaper M, Schmitt P, Nassar JE, Singh M, Farias MJ, Kim J, Diebo BG, Daniels AH. Evaluation of retrieval-augmented generation and large language models in clinical guidelines for degenerative spine conditions. Eur Spine J. 2025 Jul 7. doi: 10.1007/s00586-025-08994-8.
33. Cevik AA, Abu-Zidan FM. Utilizing AI-Powered Thematic Analysis: Methodology, Implementation, and Lessons Learned. Cureus. 2025 Jun 4;17(6):e85338. doi: 10.7759/cureus.85338.
34. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024 Dec;40(6):722-729. doi: 10.1111/edt.12965.
35. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024 Jan;57(1):108-113. doi: 10.1111/iej.13985.

36. Kim H, Bibi A, Torr P, Gal Y. Detecting LLM Hallucination Through Layer-wise Information Deficiency: Analysis of Unanswerable Questions and Ambiguous Prompts. ArXiv. 2024;abs/2412.10246. doi:10.48550/arXiv.2412.10246.

Figure Legends

Figure 1. Architecture of the Two LLM Configurations Evaluated. Google Gemini operates as a base model without document integration, generating responses solely from its pre-trained knowledge. In contrast, NotebookLM uses a retrieval-augmented generation setup, incorporating uploaded clinical references such as Heithersay's classification, Patel et al.'s reviews, and the European Society of Endodontology Position Statement. This configuration simulates the Retrieval-augmented Information System for Enhancement framework by grounding answers in embedded scientific evidence.

Figure 2. Accuracy comparison between AI models. Bar chart showing the percentage of correct responses generated by Google Gemini (GG) and NotebookLM (NLM) across 46 yes/no clinical questions related to Eternal Cervical Resorption.
Accuracy is defined as alignment with gold standard answers. NLM achieved 96% accuracy and GG 89%, but this difference was not statistically significant ($p>0.05$). Error bars represent 95% confidence intervals (Wald method).

Figure 3. Consistency Comparison Between AI Models. Bar chart illustrating the consistency of responses from Google Gemini and NotebookLM (NLM). Consistency was defined as providing identical outputs across three independent trials per question. Both models achieved 93% consistency, with no statistically significant difference ($p>0.05$). Error bars represent 95% confidence intervals.

Question No.	Question Summary	Gold Standard Answer	NLM Response	NLM Accuracy	GG Response	GG Accuracy
1	Is periapical radiography recommended for the assessment of ECR treatment?	No	No	✓	Yes	✗
3	Should Class IV ECR lesions be treated?	No	No	✓	Yes	✗
17	Is external approach preferred for large buccal portal in anterior teeth?	No	Yes	✗	Yes	✗
25	Are periapical radiographs suggested under ALARA for diagnosis?	Yes	No	✗	Yes	✓
42	Does Patel's 3D classification allow circumferential assessment?	No	Yes	✗	Yes	✓
46	Was Heithersay's success rate 100% for Class 1/2 ECR?	Yes	Yes	✓	No	✗

Table 1. Item-Level Error Analysis for Google Gemini (GG) and NotebookLM (NLM) on External Cervical Resorption Clinical Questions (ECR). This table lists the specific questions where at least one model deviated from the expert-defined gold standard. For each question, model responses, internal consistency, and final accuracy classification (accurate/inaccurate) are presented to highlight item-specific weaknesses and error patterns.

Model	Accuracy (%)	CI 95% (Wald)	CI 95% (Wilson)	Consistency (%)	Fleiss' kappa	CI 95% (Kappa)	p-value (Kappa)	p-value (Accuracy Comparison)	p-value (Consistency Comparison)
GG	89.13	[76.43, 96.37]	[76.96, 95.26]	93.48	0.89	[0.67, 1.00]	< 0.001	0.431	-
NLM	95.65	[85.16, 99.46]	[85.46, 98.79]	93.48	0.9	[0.67, 1.00]	< 0.001	-	1.000

Table 2. Overall Accuracy and Consistency of Google Gemini (GG) and NotebookLM (NLM) in Managing External Cervical Resorption-Related Questions. This table summarizes the performance of both LLMs on the full set of 46 questions. Accuracy refers to the proportion of responses that matched the expert-derived gold standard. Consistency represents the percentage of questions where each model provided identical responses across three user accounts.

Gold standard answers were derived from Heithersay's classification (10), Patel's diagnostic and treatment protocols (27,28), the European Society of Endodontology Position Statement [29], and a clinical investigation published in a Q1-ranked journal (30).

Note: Fleiss' kappa was used to evaluate inter-account agreement: 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect. Confidence intervals were calculated using the Wald and Wilson methods, where applicable. Fisher's exact test showed no significant differences between the models in either accuracy ($p = 0.431$) or consistency ($p = 1.000$), despite high agreement ($p < 0.001$) observed within each model.

Figure 1. Comparative Model Architecture

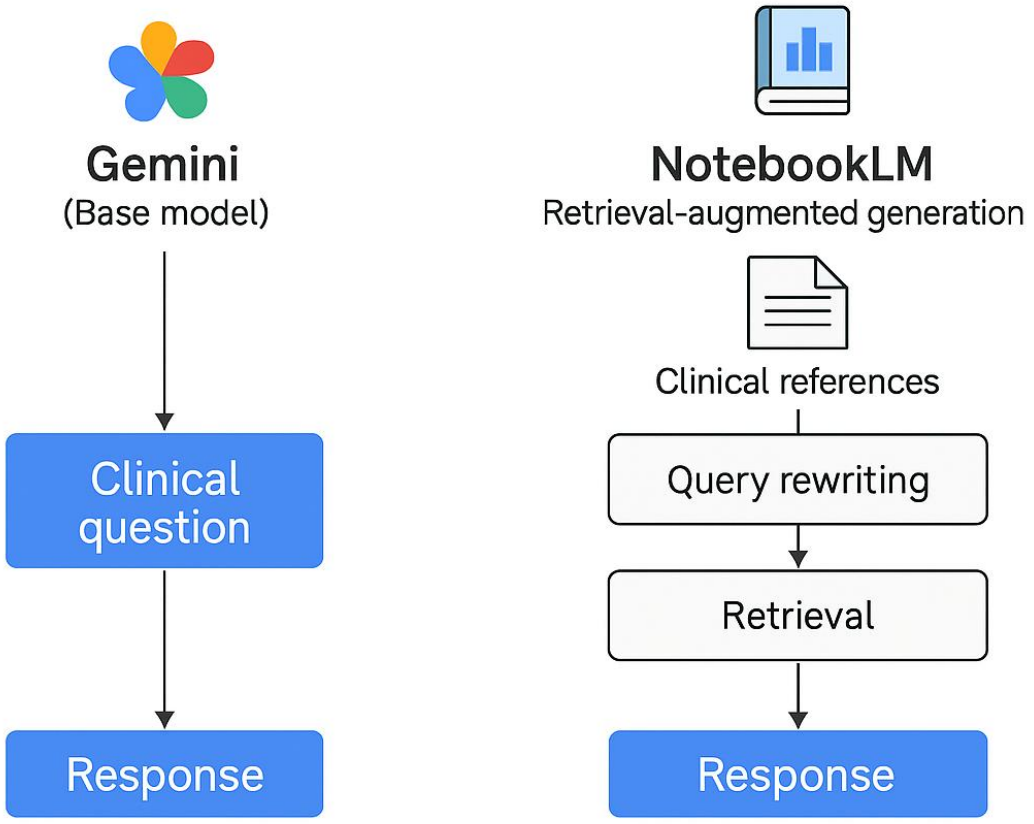

Figure 1. Architecture of the Two LLM Configurations Evaluated. Google Gemini operates as a base model without document integration, generating responses solely from its pre-trained knowledge. In contrast, NotebookLM uses a retrieval-augmented generation setup, incorporating uploaded clinical references such as Heithersay's classification, Patel et al.'s reviews, and the European Society of Endodontology Position Statement. This configuration simulates the Retrieval-augmented Information System for Enhancement framework by grounding answers in embedded scientific evidence.

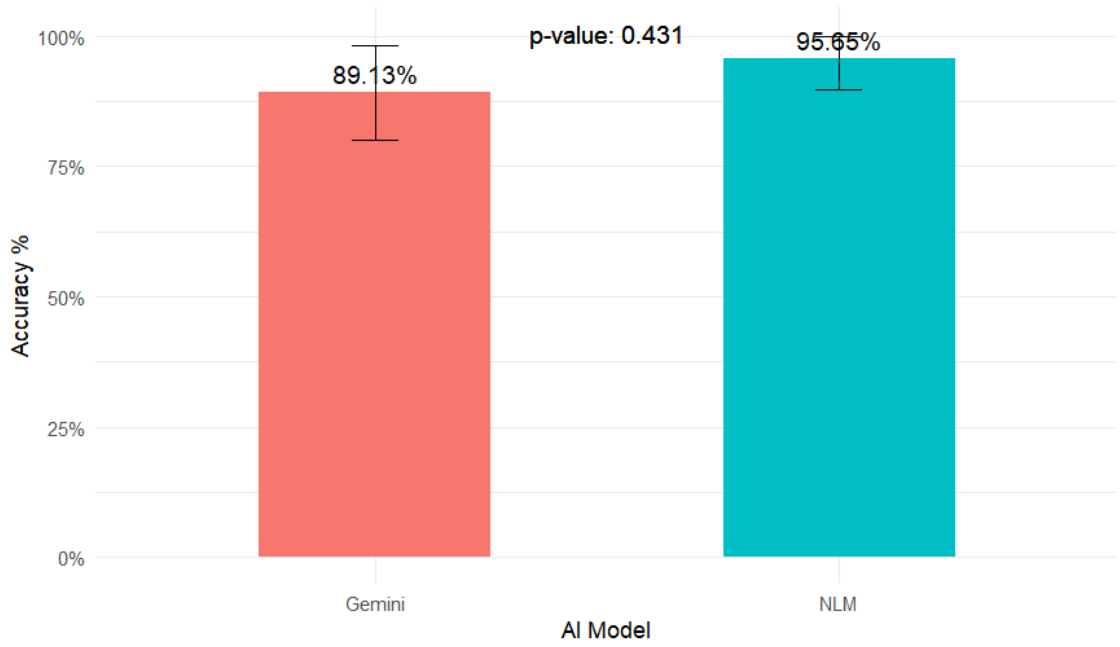

Figure 2. Accuracy comparison between AI models. Bar chart showing the percentage of correct responses generated by Google Gemini (GG) and NotebookLM (NLM) across 46 yes/no clinical questions related to Eternal Cervical Resorption.

Accuracy is defined as alignment with gold standard answers. NLM achieved 95.65% accuracy and GG 89.13%, but this difference was not statistically significant (*p = 0.431*). Error bars represent 95% confidence intervals (Wald method).

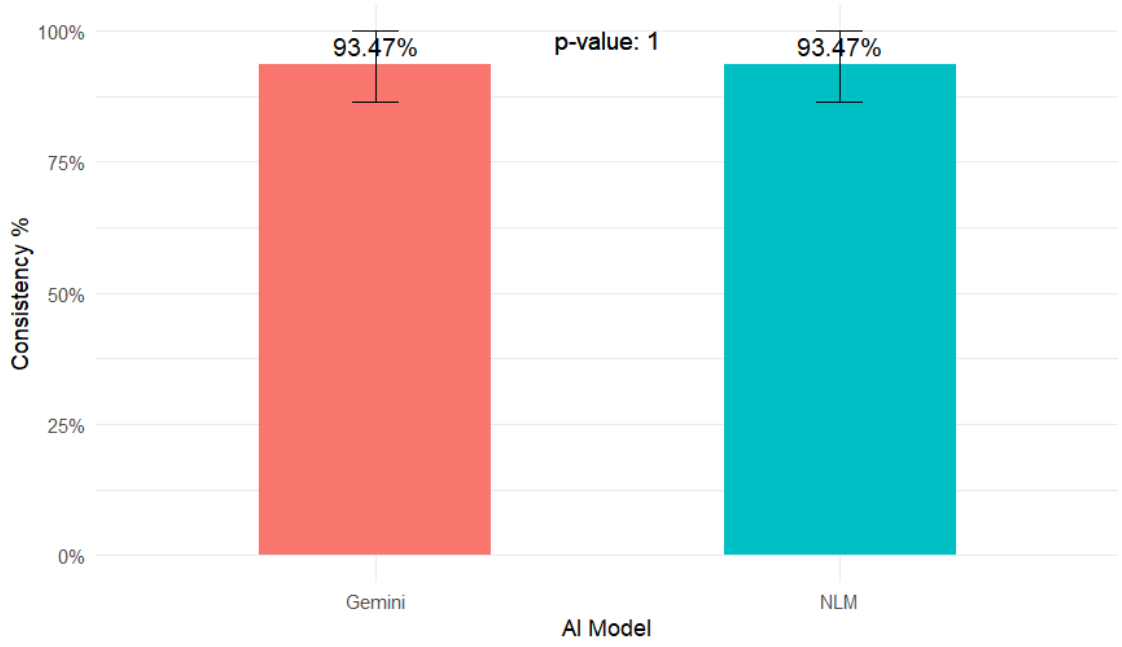

Figure 3. Consistency Comparison Between AI Models. Bar chart illustrating the consistency of responses from Google Gemini and NotebookLM (NLM). Consistency was defined as providing identical outputs across three independent trials per question. Both models achieved 93.47% consistency, with no statistically significant difference (*p = 1.000*). Error bars represent 95% confidence intervals.