

Comparative Reliability of Gemini and NotebookLM on AAPD Guideline-Based Pediatric Dentistry Multiple-Choice Questions: Accuracy, Consistency, and Error-Pattern Analysis

Junsheng Lao

Xiangya Stomatological Hospital, Central South University

Linyong Xu

linyongxu@csu.edu.cn

Xiangya Stomatological Hospital, Central South University

Research Article

Keywords: pediatric dentistry, large language models, retrieval-augmented generation, Gemini, NotebookLM

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10990696/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

Aim: To evaluate the reliability characteristics of a standalone Gemini chat workflow and a document-grounded NotebookLM workflow when answering multiple-choice pediatric dentistry questions derived from the 2024 American Academy of Pediatric Dentistry (AAPD) guideline on vital pulp therapies in primary teeth, with particular emphasis on the effects of retrieval-augmented generation (RAG) on evidence-mapping stability, error structure, and potential translational risk.

Methodology: This cross-sectional simulation study included 90 single-best-answer pediatric dentistry multiple-choice questions derived from the AAPD guideline. Each item contained four options (A-D) and a predefined gold-standard answer. Gemini and NotebookLM were each tested using three independent accounts; each account answered all 90 questions, yielding 540 responses in total. The standardized prompt was, "You are an experienced physician," and a new conversation was initiated for each query. For NotebookLM, the 2024 AAPD vital pulp therapy guideline PDF was uploaded as a fixed document source. The primary outcome was majority-vote accuracy, and the secondary outcome was inter-account consistency. In addition to Wald and Wilson 95% confidence intervals, Fleiss' kappa, Pearson chi-squared tests, and Fisher's exact tests, an error-structure framework was developed to classify errors as stable or fluctuating and to examine evidence-to-option mapping failures.

Results: Gemini achieved a majority-vote accuracy of 92.22% (83/90; Wald 95% CI: 86.69%-97.76%; Wilson 95% CI: 84.81%-96.18%) and an inter-account consistency of 83.33% (75/90; Wilson 95% CI: 74.31%-89.63%; Fleiss' kappa = 0.829). NotebookLM achieved a majority-vote accuracy of 98.89% (89/90; Wald 95% CI: 96.72%-100.00%; Wilson 95% CI: 93.97%-99.80%) and an inter-account consistency of 100.00% (90/90; Wilson 95% CI: 95.91%-100.00%; Fleiss' kappa = 1.000). For accuracy, Pearson's chi-squared test indicated a statistically significant difference ($p = 0.030$), whereas Fisher's exact test did not reach the 0.05 threshold ($p = 0.064$), suggesting that the apparent accuracy advantage should be interpreted cautiously. For consistency, NotebookLM was significantly higher in both tests (both $p < 0.001$). Gemini produced seven majority-vote errors, including one stable error and six fluctuating errors. NotebookLM produced only one error (Q37), which was stable. Errors were concentrated in items requiring precise evidence mapping, including comparisons of caries removal techniques, quantitative outcomes, treatment success rates, Hall technique effects, and diagnostic differentiation.

Conclusions: In this pediatric dentistry guideline-based multiple-choice benchmark, the clearest advantage of RAG was not simply an increase in the number of correct answers, but rather improved cross-account stability and reduced random drift. Nevertheless, document grounding did not fully eliminate evidence-to-option mapping failures. Because stable errors are consistent and may appear evidence-based, they may pose more concealed educational and clinical translational risks. Future evaluations of AI systems in dentistry should move beyond accuracy alone toward an integrated reliability framework encompassing correctness, consistency, error architecture, evidence-mapping behavior, and potential risk.

1 Introduction

Clinical decision-making in pediatric dentistry depends heavily on detailed conditional reasoning. The selection of treatment for deep caries and pulp therapy in primary teeth often requires integration of pain history, pulp status, radiographic findings, risk of pulp exposure, material characteristics, restorative approach, treatment success rate, and follow-up interval. ⁽¹⁾The 2024 AAPD guideline, "Use of Vital Pulp Therapies in Primary Teeth," focuses on vital pulp therapies for deeply carious primary teeth with normal pulp or reversible pulpitis, and addresses key issues such as indirect pulp treatment, direct pulp capping, pulpotomy, the Hall technique, caries removal strategies, and material selection. ⁽²⁾ Rapid, accurate, and stable translation of guideline evidence into actionable answers is therefore a practical need in pediatric dentistry education, examination training, and clinical knowledge retrieval. ⁽³⁾

Large language models (LLMs), owing to their natural language understanding and rapid text-generation capabilities, are increasingly being applied to medical knowledge question answering, case interpretation, educational support, and clinical guideline retrieval. ^(4, 5) However, in specialized medical settings, general-purpose LLMs continue to face challenges related to outdated knowledge, limited source traceability, hallucinated content, and insufficient response stability. ^(6, 7) Retrieval-augmented generation (RAG), by introducing external authoritative knowledge sources before answer generation, provides an important pathway for improving specialty alignment, evidence traceability, and output stability. ⁽⁸⁾ Prior dental studies have shown that document-grounded or RAG workflows can achieve high accuracy and consistency in guideline-oriented tasks such as dental trauma and external cervical resorption, although their advantages do not always manifest as statistically significant improvements in accuracy. ^(9, 10)

For guideline-based tasks, therefore, reporting overall accuracy alone is insufficient. ⁽¹¹⁾ Pediatric dentistry multiple-choice questions are not merely factual recall tasks; rather, they constitute structured evidence-mapping tasks. The model must first identify the clinical conditions embedded in the stem, locate the relevant guideline evidence, interpret comparative or quantitative outcomes, and finally map that evidence to the single best answer among options A-D. ⁽¹²⁾ Errors may occur at multiple stages, including recognition of stem conditions, localization of guideline evidence, interpretation of evidence, translation of quantitative outcomes, and exclusion of distractors. In other words, a model's familiarity with a treatment term does not necessarily mean that it can consistently select the correct answer under complex conditional constraints. ⁽¹¹⁾

The central question of this study was not simply whether Gemini or NotebookLM is "stronger," but whether RAG changes the reliability structure of pediatric dentistry guideline-based question answering. Specifically, RAG may reduce random outputs and cross-account variability, yet it may also generate low-frequency, reproducible, and seemingly more credible stable errors. ⁽¹³⁾ In educational and clinical support contexts, stable and fluctuating errors carry different risk implications: fluctuating errors reveal

generative uncertainty and may be more readily detected through repeated querying, whereas stable errors may be more easily mistaken for reliable answers because of their high consistency.^(14, 15)

On this basis, the present study introduced an error-structure analysis framework in addition to conventional accuracy and consistency metrics. By distinguishing stable from fluctuating errors and integrating thematic domains and knowledge attributes of incorrectly answered items, this study proposes a reliability-evaluation approach for AI question answering in pediatric dentistry guidelines. This approach simultaneously addresses whether the model is correct, whether the answer is stable, how errors arise, which types of knowledge are most affected, and what those errors may imply for education or clinical translation.

2 Materials and Methods

2.1 Ethics Statement

This study analyzed AI-generated responses and guideline-derived questions and did not involve patients, personally identifiable information, or clinical interventions.

2.2 Study Design

This was a cross-sectional simulation study using a pediatric dentistry multiple-choice question dataset with predefined gold-standard answers. It compared the performance of a standalone Gemini chat workflow with that of a document-grounded NotebookLM workflow on AAPD guideline-derived questions. The study objects were AI-generated responses and did not involve patients, personally identifiable information, or clinical interventions.

2.3 Question Selection, Thematic Domains, and Gold Standard

The question bank comprised 90 single-best-answer pediatric dentistry multiple-choice questions, all developed from knowledge points in the 2024 AAPD guideline "Use of Vital Pulp Therapies in Primary Teeth." (2) Each item included four options (A-D) and one gold-standard answer. The questions were developed by three clinicians according to the guideline and were presented in the format of a clinical knowledge assessment. The stems did not explicitly include phrases such as "according to the guideline" or "the guideline recommends," in order to approximate real examination and educational retrieval settings. The three clinicians jointly determined the gold-standard answers based on the guideline.

The 90 questions were divided into six thematic domains, with 15 items in each domain: diagnosis of pulp status and treatment indications (Q1-Q15), pulpotomy materials and medicament selection (Q16-Q30), indirect pulp treatment and caries removal techniques (Q31-Q45), direct pulp capping and treatment of special tooth sites (Q46-Q60), the Hall technique and management of special

circumstances (Q61-Q75), and operative details and prognosis (Q76-Q90). The complete questions, options, and gold-standard answers are presented as supplementary material.

2.4 Platform Specifications

Gemini testing was performed using the Gemini 3.0 Flash model via the Google Gemini web platform. NotebookLM was accessed using the same Gemini accounts, and the 2024 AAPD vital pulp therapy guideline PDF was uploaded as a fixed document-grounding source. Therefore, NotebookLM results should be interpreted as the performance of a document-grounded workflow rather than as a direct assessment of the intrinsic capability of an independent foundation model.

2.5 Response Generation Protocol

Each question was submitted separately to Gemini and NotebookLM, generating three response sets (A, B, and C). Both workflows were tested using three different Gemini accounts; each account answered all 90 questions. Thus, Gemini generated 270 responses and NotebookLM generated 270 responses, for a total of 540 responses. The standardized prompt was, "You are an experienced physician." To reduce contextual carry-over, chained memory, and influence from preceding questions, a new conversation was created for each question.

2.6 Accuracy and Consistency Assessment

Majority-vote accuracy was defined as the proportion of items for which the option appearing most frequently across the three responses from the same model matched the gold-standard answer. When the three responses were not identical, the 2/3 majority answer was used as the representative answer for that model on that item. Inter-account consistency was defined as the proportion of items for which the A, B, and C responses from the same model were identical.

To further examine error mechanisms, majority-vote errors were classified into two categories. A stable error was defined as an item for which all three responses were identical but inconsistent with the gold standard, representing a reproducible systematic deviation. A fluctuating error was defined as an item for which the three responses were not identical and the majority-vote answer was incorrect, representing instability or option drift during generation. This classification was used to describe error structure and did not replace expert clinical review of the question content itself.

2.7 Statistical Analysis

All data were analyzed in R. Accuracy and consistency were expressed as proportions, and Wald and Wilson 95% confidence intervals were calculated for each. Inter-account agreement was quantified using Fleiss' kappa. Differences in accuracy and consistency between the two workflows were assessed using Pearson's chi-squared test, with Fisher's exact test used as a sensitivity analysis. Because NotebookLM produced few errors, Fisher's exact test was particularly important for interpreting differences in accuracy. The threshold for statistical significance was set at $p < 0.05$.

3 Results

3.1 Gemini

Gemini answered 83 items correctly by majority vote, corresponding to an overall accuracy of 92.22% (83/90), an inter-account consistency of 83.33% (75/90), and Fleiss' kappa of 0.829. Its majority-vote error items were Q14, Q35, Q37, Q50, Q52, Q62, and Q73.

3.2 NotebookLM

NotebookLM answered 89 items correctly by majority vote, corresponding to an overall accuracy of 98.89% (89/90), an inter-account consistency of 100.00% (90/90), and Fleiss' kappa of 1.000. Its only majority-vote error was Q37, for which all three responses were identical.

3.3 Comparison Between Workflows

A total of 90 pediatric dentistry multiple-choice questions and 540 model responses were analyzed. Gemini answered 83 items correctly by majority vote, whereas NotebookLM answered 89 items correctly. Gemini produced identical responses across the three accounts for 75 items, whereas NotebookLM produced identical responses for all 90 items. Overall results are shown in Table 1.

Table 1
Overall performance of Gemini and NotebookLM on 90 AAPD guideline-derived pediatric dentistry multiple-choice questions.

Metric	Gemini	NotebookLM
Majority-vote accuracy	92.22% (83/90)	98.89% (89/90)
Accuracy Wald 95% CI	86.69%-97.76%	96.72%-100.00%
Accuracy Wilson 95% CI	84.81%-96.18%	93.97%-99.80%
Inter-account consistency	83.33% (75/90)	100.00% (90/90)
Consistency Wald 95% CI	75.63%-91.03%	100.00%-100.00%
Consistency Wilson 95% CI	74.31%-89.63%	95.91%-100.00%
Fleiss' kappa	0.829	1.000

In the accuracy comparison, NotebookLM had a higher point estimate than Gemini (98.89% vs. 92.22%). Pearson's chi-squared test suggested a statistically significant difference ($p = 0.030$), whereas Fisher's exact test did not reach the conventional significance threshold ($p = 0.064$). Given that NotebookLM had only one incorrect item and the number of events was small, the accuracy difference should be interpreted cautiously and exploratorily. By contrast, the difference in consistency was more robust: NotebookLM achieved 100.00% consistency compared with 83.33% for Gemini, with both Pearson's chi-squared test and Fisher's exact test yielding $p < 0.001$. The difference in majority-vote accuracy between

the two workflows is illustrated in Fig. 1. The comparison of inter-account consistency between the two workflows is shown in Fig. 2.

3.4 Item-Level Analysis and Error Distribution

Gemini's majority-vote error items were Q14, Q35, Q37, Q50, Q52, Q62, and Q73. Among these, Q35 was a stable error, whereas Q14, Q37, Q50, Q52, Q62, and Q73 were fluctuating errors. NotebookLM produced a majority-vote error only on Q37, and all three responses were identical, making it a stable error. The error structure is shown in Table 2.

Table 2
Majority-vote error items and error-structure classification.

Model	Majority-vote error items	Stable error	Fluctuating error	Error-structure characteristics
Gemini	Q14, Q35, Q37, Q50, Q52, Q62, Q73	Q35	Q14, Q37, Q50, Q52, Q62, Q73	Predominantly cross-account variability and option drift
NotebookLM	Q37	Q37	None	Low-frequency but reproducible stable error

By thematic domain, Gemini's errors were not uniformly distributed. One error occurred in diagnosis of pulp status and treatment indications (Q14), two in indirect pulp treatment and caries removal techniques (Q35 and Q37), two in direct pulp capping and treatment of special tooth sites (Q50 and Q52), and two in the Hall technique and management of special circumstances (Q62 and Q73). No majority-vote errors occurred in the domains of pulpotomy materials and medicament selection or operative details and prognosis. NotebookLM's only incorrect item, Q37, was in the domain of indirect pulp treatment and caries removal techniques. The distribution of error items across the six thematic domains is summarized in Table 3. The item-level distribution of majority-vote errors, including stable and fluctuating errors, is presented in Fig. 3.

Table 3
Distribution of error items across the six thematic domains.

Thematic domain	Question range	Gemini error count	NotebookLM error count	Error items
Diagnosis of pulp status and treatment indications	Q1-Q15	1/15	0/15	Q14
Pulpotomy materials and medicament selection	Q16-Q30	0/15	0/15	None
Indirect pulp treatment and caries removal techniques	Q31-Q45	2/15	1/15	Q35, Q37
Direct pulp capping and treatment of special tooth sites	Q46-Q60	2/15	0/15	Q50, Q52
Hall technique and management of special circumstances	Q61-Q75	2/15	0/15	Q62, Q73
Operative details and prognosis	Q76-Q90	0/15	0/15	None

In terms of knowledge attributes, the incorrect items mainly involved comparative or quantitative information. Q35 concerned the incidence of pulp exposure associated with complete caries removal; Q37 involved comparison of chemomechanical caries removal and rotary-bur caries removal; Q50 involved the number needed to treat (NNT); Q52 concerned the success rate of vital pulp therapy in primary molars with proximal caries; Q62 addressed the effect of the Hall technique on maintaining pulp vitality; and Q73 involved comparison of 24-month success rates between indirect pulp treatment with a preformed crown and pulpotomy. Q14 was a diagnostic differentiation item requiring identification of reliable criteria for distinguishing pulp necrosis from irreversible pulpitis in primary teeth. This distribution suggests that models performed more stably on concept-recognition items, while errors were more likely in items involving evidence comparison, conditional constraints, and quantitative outcomes.

To further explain the reliability of model outputs, each item was placed into a two-dimensional correctness-stability framework. This framework considers not only whether an answer is correct, but also whether the model produces the same output consistently across accounts. Under this framework, AI responses can be categorized as stable correct, fluctuating correct, fluctuating error, or stable error. Under this framework, AI responses can be categorized as stable correct, fluctuating correct, fluctuating error, or stable error. The definitions, risk interpretations, and meanings of these four response types in the present study are summarized in Table 4.

Table 4
Error-architecture framework for pediatric dentistry guideline-based AI question answering.

Type	Definition	Risk interpretation	Meaning in this study
Stable correct	Three account responses are identical and the majority-vote answer is correct	Low risk	Indicates that the model matches the gold standard and has cross-account stability
Fluctuating correct	Three account responses are not identical, but the majority-vote answer is correct	Moderate risk	Indicates that the model may answer correctly but shows generative uncertainty
Fluctuating error	Three account responses are not identical and the majority-vote answer is incorrect	High risk	Suggests drift among similar options, so repeated use may produce different recommendations
Stable error	Three account responses are identical but inconsistent with the gold standard	Concealed high risk	The error is consistent and authoritative in appearance, making it harder for users to detect

4 Discussion

This study found that both Gemini and NotebookLM achieved high majority-vote accuracy on AAPD guideline-derived pediatric dentistry multiple-choice questions.⁽²⁾ If overall accuracy alone is considered, both workflows could be described as performing well. However, when consistency and error structure are incorporated, the two workflows show clearly distinct reliability profiles. Gemini's main limitation was cross-account variability: six of its seven majority-vote errors were fluctuating errors. NotebookLM's defining feature was high output stability, with complete inter-account consistency across all 90 items, although it still produced one stable error.

These findings indicate that the value of RAG should not be reduced to "improving accuracy." In this dataset, NotebookLM had a higher point estimate for accuracy than Gemini, but Fisher's exact test did not reach the 0.05 threshold, indicating that the accuracy advantage requires validation in larger samples. By contrast, the consistency advantage was significant in both statistical tests, suggesting that the more clearly demonstrable effect of the document-grounded workflow was to reduce random drift and improve cross-account reproducibility. In other words, in this task, RAG primarily changed the stability structure of model outputs rather than necessarily producing a statistically robust improvement in accuracy.⁽¹⁶⁾

The content of the incorrect items suggests that model errors occurred mainly at the evidence-to-option mapping stage rather than at the level of simple factual recall.⁽¹⁷⁾ Pediatric dentistry guideline-based multiple-choice questions commonly include several closely related options and require the model to

select the single best answer after interpreting guideline evidence. For items involving treatment success rates, NNT, risk of pulp exposure, Hall technique effects, and comparisons among treatment strategies, the model must simultaneously process clinical conditions, comparators, outcome measures, and time horizons. Inadequate interpretation at any of these stages may lead to an incorrect final option. (18, 19)

Evidence-to-option mapping failure can therefore be regarded as a core error mechanism in guideline-based AI question answering. ^(20, 21)It differs from hallucination in the conventional sense, because the model may not be fabricating nonexistent information. It also differs from simple knowledge absence, because the model may have access to or have internalized the relevant knowledge, yet still deviate at the stages of evidence prioritization, applicability assessment, or option matching. ⁽²²⁾This finding suggests that RAG can improve evidence availability but does not necessarily guarantee correct evidence interpretation or option mapping.

Stable and fluctuating errors represent different risk patterns. Fluctuating errors reflect inconsistent outputs for the same item across accounts or sessions, suggesting uncertainty in the model's internal representation or generation process. In educational contexts, fluctuating errors may lead students to receive different answers at different times, thereby reducing trust in learning support; however, such inconsistency may also prompt users to perform secondary verification. ⁽¹⁵⁾

Stable errors are more concealed. NotebookLM's only error, Q37, was identical across the three accounts, indicating that document grounding did not prevent a systematic deviation for that item. Stable errors often have a stronger appearance of credibility: the answer is consistent, the tone is confident, and the document-grounded context may lead users to regard the output as evidence-based. ⁽²³⁾If users do not actively check the original guideline, stable errors may be more likely to be accepted as reliable. Safety evaluation of RAG systems should therefore consider not only the number of errors but also whether errors are reproducible, whether they involve key treatment decisions, and whether they carry a misleading appearance of authority. ^(21, 23)

Previous dental trauma research comparing a standalone Gemini chat workflow with document-grounded NotebookLM found that NotebookLM achieved higher accuracy and complete consistency on guideline-derived yes/no questions. A study on external cervical resorption also compared Gemini and NotebookLM but found no significant differences in accuracy or consistency. ^(9, 10)Together, these findings suggest that the benefits of RAG are task-dependent. When a task depends heavily on explicit guideline statements and reproducible outputs, RAG is more likely to manifest as a consistency advantage; when the foundation model already covers the task knowledge well, or when the items are closed binary questions, the marginal gain in accuracy from RAG may be smaller. ⁽²⁴⁾

The present study extends this issue to pediatric dentistry multiple-choice questions. Compared with yes/no questions, A-D multiple-choice questions introduce additional requirements for distractor exclusion and option mapping, making them more likely to reveal model weaknesses in evidence

interpretation, quantitative outcome processing, and comparative reasoning.⁽²⁵⁾ The present findings are consistent with the broader trend in dental RAG research: document grounding can improve reproducibility and task alignment, but it cannot fully eliminate evidence misinterpretation, conditional mismatch, or systematic error.

For pediatric dentistry education, this study suggests that LLMs may serve as auxiliary tools for guideline learning and knowledge self-assessment, but they should not be treated as answer arbiters that require no verification.^(26, 27) Educators may use fluctuating errors to identify knowledge boundaries that are easily confused by both students and models, such as comparisons of caries removal techniques, Hall technique effects, and treatment success rates. Stable errors should also be treated as key discussion points, guiding students back to the original guideline to understand the relationships among evidence, outcome measures, and applicability conditions.

For clinical retrieval support, this study emphasizes that guideline concordance is not equivalent to clinical competence. A multiple-choice benchmark can evaluate the extent to which a model aligns with guideline answers under controlled conditions, but real clinical decision-making additionally requires individualized risk assessment, uncertainty management, patient preferences, operative conditions, follow-up feasibility, and other multidimensional considerations.⁽²⁸⁾ Especially when model outputs involve treatment selection, prognostic judgment, NNT, success rates, or critical thresholds, traceable evidence should be required and independent professional verification should be performed.

The methodological significance of this study lies in shifting AI evaluation from simple performance comparison toward reliability diagnosis. Traditional benchmarks usually answer only how many items a model answers correctly. In medical education and clinical support contexts, however, more important questions include whether the model is stable on the same item, whether errors are reproducible, which knowledge types are most error-prone, whether errors involve high-risk treatment decisions, and whether users are likely to detect those errors.^(29, 30)

The proposed framework of accuracy, consistency, error type, item-content characteristics, and translational risk may therefore serve as a basis for future evaluations of guideline-based AI systems in pediatric dentistry. This framework is applicable not only to the AAPD vital pulp therapy guideline but may also be extended to other guideline-based tasks, including pediatric dental trauma, caries management, sedation and analgesia, and preventive oral health care. Future studies may further combine this framework with case-based vignettes, open-ended question answering, verification of original-source citations, and error-severity grading to better approximate real clinical and educational settings.

4.7 Limitations

This study has several limitations. First, although the 90-item question set allowed preliminary statistical comparison, the study may still have been underpowered for low-frequency errors and small differences

in accuracy. Second, the multiple-choice format ensured objective scoring and reproducibility, but it could not fully simulate the openness, continuity, and individualized judgment involved in real clinical reasoning. Third, NotebookLM was evaluated as a document-grounded workflow rather than as an independent foundation model; therefore, the results should be interpreted as workflow-level performance available to users and should not be directly extrapolated to differences in underlying model capability. Fourth, the content analysis of incorrect items was based on the current question bank and gold standard; future work should include further expert review of shared error items to distinguish item discrimination, guideline evidence interpretation, and model mapping bias.

4.8 Future Research Directions

Future research should be deepened in three directions. First, the question bank should be expanded and repeated testing across time points should be added to examine the effects of model updates, prompt changes, and temporal drift on stability. Second, case-based vignettes and open-ended explanatory questions should be introduced to evaluate conditional reasoning, evidence citation, and uncertainty expression across realistic pediatric dentistry decision chains. Third, evidence-verification mechanisms should be established to determine not only whether the model provides the correct answer, but also whether the guideline evidence it cites genuinely supports the conclusion. For stable errors, error-item tracking and mandatory expert review should be implemented; for fluctuating errors, prompt optimization, structured output, and majority-vote strategies may be used to reduce random drift.

5 Conclusions

Based on testing with 90 AAPD guideline-derived pediatric dentistry multiple-choice questions, both Gemini and NotebookLM demonstrated high majority-vote accuracy. NotebookLM had a higher point estimate for accuracy, although Fisher's exact test indicated that this finding should be interpreted cautiously. Its clearest advantage was an inter-account consistency of 100%, which was significantly higher than that of Gemini.

More importantly, the two workflows exhibited different error architectures. Gemini's main limitation was fluctuating error, reflecting unstable generation across accounts; NotebookLM's main risk was low-frequency but stable error, indicating that document grounding cannot fully eliminate evidence-mapping failure. Model errors were concentrated in items requiring precise evidence mapping, including comparisons of treatment techniques, quantitative outcomes, success-rate judgments, Hall technique effects, and diagnostic differentiation.

Therefore, this study should not be understood merely as a horizontal performance comparison between Gemini and NotebookLM. Rather, it should be viewed as an exploration of a reliability-evaluation paradigm for pediatric dentistry guideline-based AI systems. Future research and applications should simultaneously consider whether a model is correct, whether it is stable, why it makes errors, what knowledge categories those errors involve, and what risks such errors may pose in educational or clinical translation.

Declarations

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval

This study analyzed AI-generated responses to guideline-derived multiple-choice questions and did not involve human participants, patients, human tissue, personally identifiable information, or clinical interventions. Therefore, formal ethics approval was not required. The relevant institutional ethics body is the Ethics Committee of Xiangya Stomatological Hospital and Xiangya School of Stomatology, Central South University.

Human Ethics and Consent to Participate

Human Ethics and Consent to Participate declarations: not applicable.

Consent to Participate

Not applicable. This study did not involve human participants.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Dhar V, El-Karim I, Coll JA, Fouad AF, O'Connell AC, Asgary S, et al. Consensus-based recommendations on pulp therapies in primary and permanent teeth: iapd porto forum. *Int J Paediatr Dent*. 2026;36(2):371–86. 10.1111/ipd.70068.
2. Coll JA, Dhar V, Chen C, Crystal YO, Guelmann M, Marghalani AA, et al. Use of vital pulp therapies in primary teeth 2024. *Pediatr Dent*. 2024;46(1):13–26.
3. Bardakci E, Aydin G, Dogan ME, Dogan MS. Ai-based decision support in pediatric dentistry: a comparative study of chatgpt-5 and gemini advanced. *BMC Oral Health*. 2026;26(1). 10.1186/s12903-026-07897-4.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–40. 10.1038/s41591-023-02448-8.
5. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. *Bioeng (Basel)*. 2025;12(6). 10.3390/bioengineering12060631.

6. Fernandes AC, Souto MEVC. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *N Engl J Med*. 2023;388(25):2399–400. 10.1056/NEJMc2305286.
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–80. 10.1038/s41586-023-06291-2.
8. Ghosh S, Mittal G. Advancing engineering research through context-aware and knowledge graph-based retrieval-augmented generation. *Front Artif Intell*. 2025;8:1697169. 10.3389/frai.2025.1697169.
9. Dufey-Portilla N, Abella Sans F, Duran-Sindreu F, Salagaray MT, Nagendrababu V, Abbott PV, et al. Evaluating guideline adherence in gemini-powered dental trauma workflows: standalone gemini chat vs. Document-grounded notebooklm. *Dent Traumatol*. 2026. 10.1111/edt.70065.
10. Garcia-Font M, Dufey-Portilla N, Durán-Sindreu F, González Sánchez JA, Rodríguez Millán G, Nagendrababu V, et al. Evaluating retrieval-augmented large language models on external cervical resorption: a comparative study of gemini and notebooklm. *J Endod*. 2026;52(2):300–06. 10.1016/j.joen.2025.10.016.
11. Mukhopadhyay A, Mukhopadhyay S, Biswas R. Evaluation of large language models in pediatric dentistry: a bloom's taxonomy-based analysis. *Folia Med (Plovdiv)*. 2025;67(4). 10.3897/folmed.67.e154338.
12. Kocaoğlu MH, Demirel A, Kaya İ. Accuracy, quality, and readability analyses of responses from large language models to questions on pediatric dental sedation. *BMC Oral Health*. 2026;26(1). 10.1186/s12903-026-08026-x.
13. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. *Plos Digit Health*. 2025;4(6):e877. 10.1371/journal.pdig.0000877.
14. Gürses ÖA, Ceylan İ. Consistency over accuracy: run-to-run stability of contemporary large language models on turkish curriculum-aligned theoretical anatomy multiple-choice questions. *Bmc Med Educ*. 2026;26(1):285. 10.1186/s12909-026-08656-3.
15. Funk PF, Hoch CC, Knoedler S, Knoedler L, Cotofana S, Sofo G, et al. Chatgpt's response consistency: a study on repeated queries of medical examination questions. *Eur J Investig Health Psychol Educ*. 2024;14(3):657–68. 10.3390/ejihpe14030043.
16. Yao X, Nédellec C, Xia J, Bossy R. Consistency-accuracy correlation in hard-prompted llms for entity and relation extraction: empirical findings from plant-health data. *Genomics Inf*. 2026;24(1):3. 10.1186/s44342-025-00063-2.
17. Strasser LM, Anschuetz W, Dennstädt F, Hastings J. Performance evaluation of large language models in multilingual medical multiple-choice questions: mixed methods study. *Jmir Med Educ*. 2026;12:e81399. 10.2196/81399.
18. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nat Commun*. 2025;16(1):9799. 10.1038/s41467-025-64769-1.

19. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (N Y)*. 2024;5(3):100943. 10.1016/j.patter.2024.100943.
20. Wang C, Chen Y. Evaluating large language models for evidence-based clinical question answering. *Patterns (N Y)*. 2026;7(5):101519. 10.1016/j.patter.2026.101519.
21. Masannek L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented llms with document or online support for evidence based neurology. *Npj Digit Med*. 2025;8(1):137. 10.1038/s41746-025-01536-y.
22. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. *Nat Commun*. 2025;16(1):642. 10.1038/s41467-024-55628-6.
23. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *Npj Digit Med*. 2025;8(1):274. 10.1038/s41746-025-01670-7.
24. Giuffrè M. Distinguishing retrieval augmented generation from prompt engineering: implications for reproducibility in large language model research and applications. *Am J Gastroenterol*. 2025;120(11):2727. 10.14309/ajg.0000000000003439.
25. Diniz P, Yokoe T, Öttl FC, Pereira H, Henriques R, Samuelsson K. Reasoning-optimised large language models reach near-expert accuracy on board-style orthopaedic exams: a multi-model comparison on 702 multiple-choice questions. *Knee Surg Sports Traumatol Arthrosc*. 2026;34(2):752–62. 10.1002/ksa.70222.
26. Akitomo T, Hamada M, Hamaguchi S, Kusaka S, Nomura R. Generative artificial intelligence and large language models in paediatric dentistry: a scoping review. *Int Dent J*. 2026;76(2):109401. 10.1016/j.identj.2025.109401.
27. Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S. Can large language models serve as reliable tools for information in dentistry? A systematic review. *Int Dent J*. 2025;75(4):100835. 10.1016/j.identj.2025.04.015.
28. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof*. 2004;27(3):304–19. 10.1177/0163278704267043.
29. Wang Q, Gong Z, Lai Z, Bu L, Dahlweid F, Sun H. Promises and challenges of applying large language models in the healthcare domain. *Front Digit Health*. 2026;8:1772274. 10.3389/fdgth.2026.1772274.
30. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Freeman R, Nadkarni GN, et al. Challenges of implementing llms in clinical practice: perspectives. *J Clin Med*. 2025;14(17). 10.3390/jcm14176169.

Figures

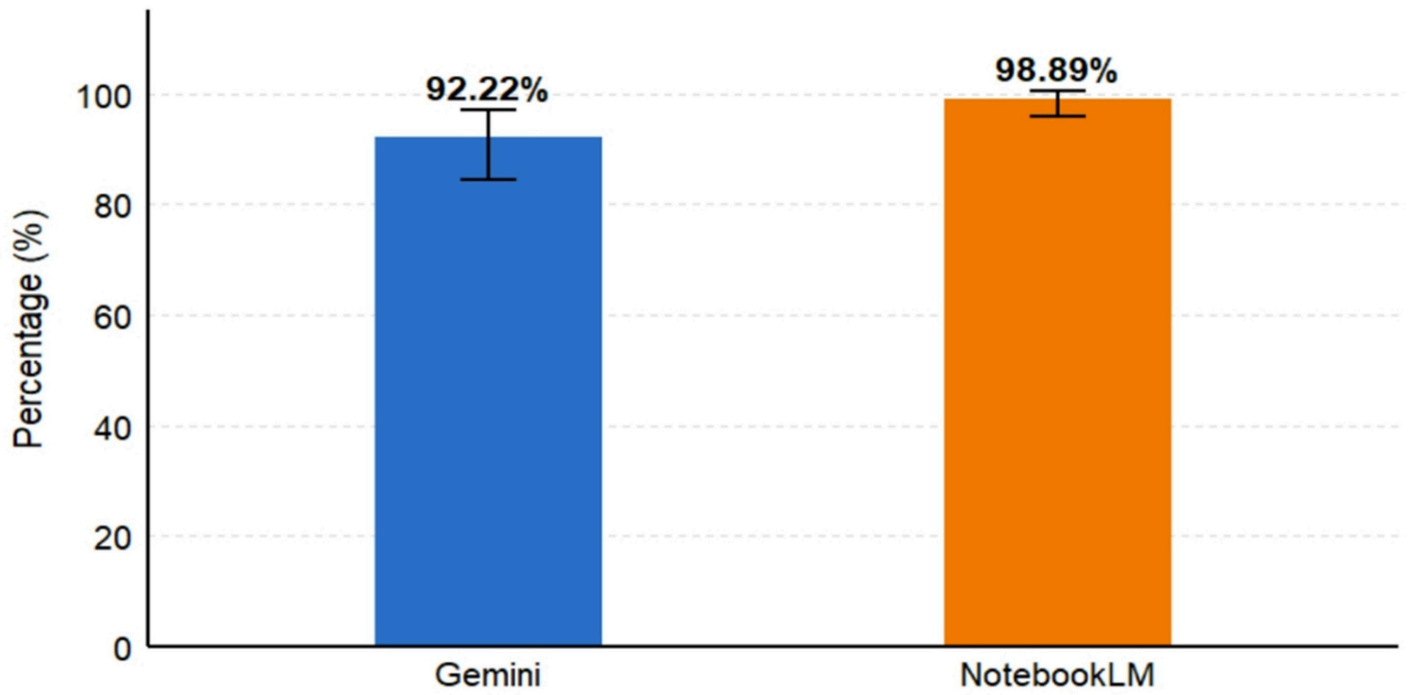


Figure 1

Majority-vote accuracy of Gemini and NotebookLM. Error bars indicate Wald 95% confidence intervals.

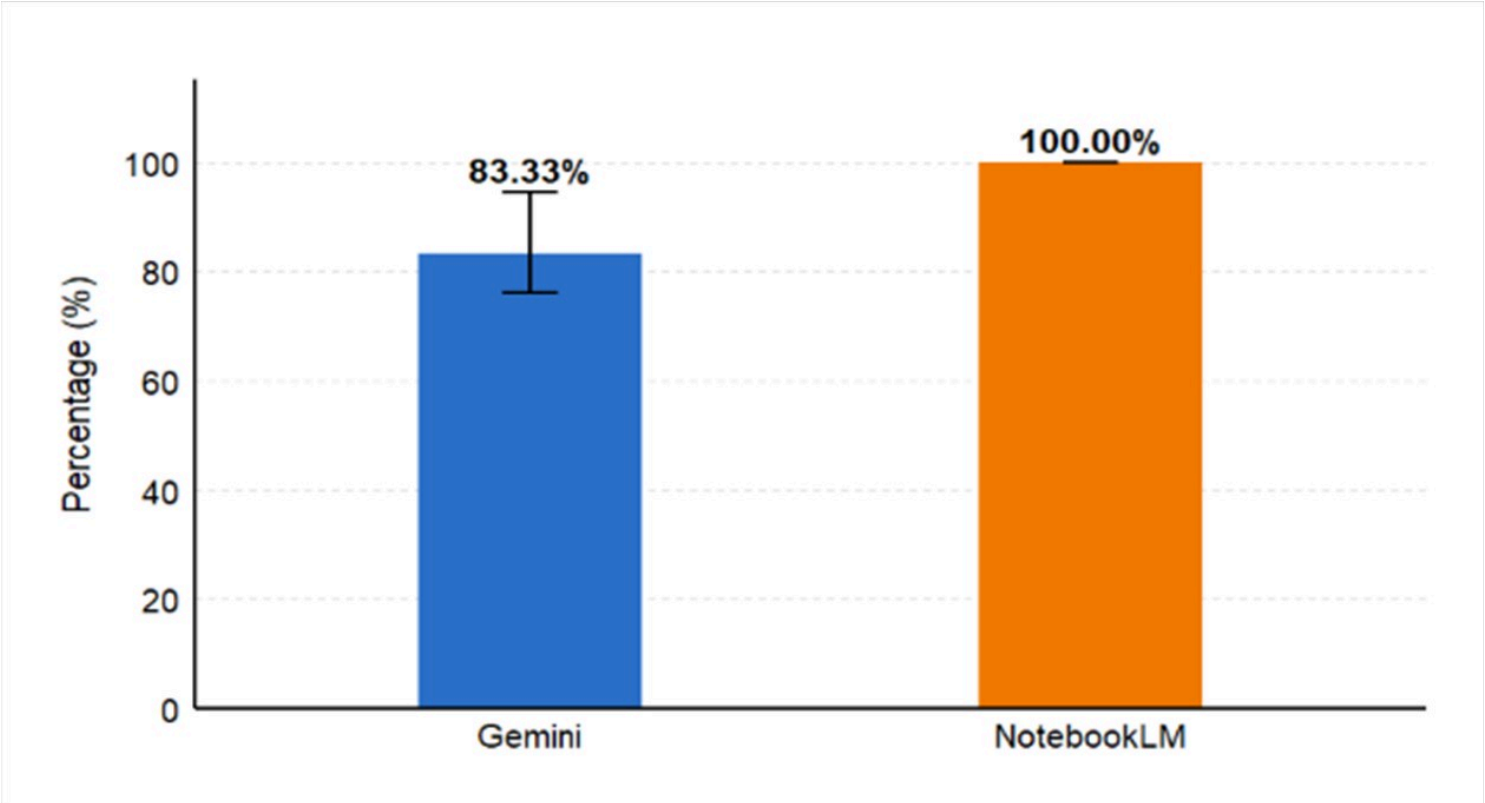


Figure 2

Inter-account consistency of Gemini and NotebookLM. Error bars indicate Wald 95% confidence intervals.



Figure 3

Item-level distribution of majority-vote error items. Circles indicate stable errors and diamonds indicate fluctuating errors.

Supplementary Files

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

- [DataLogbook.xlsx](#)
- [TestQuestionsandGoldStandardAnswers.docx](#)