



Evaluation of the accuracy and robustness of large language models in dental examinations: sensitivity analysis of question format and evaluation parameters

Evaluación de la precisión y robustez de los modelos de lenguaje de gran tamaño en exámenes de odontología: análisis de sensibilidad del formato de las preguntas y los parámetros de evaluación

Andrea Atassi¹
Shachar Stewart¹
Carrie Brackenridge¹
Eliza Kamin¹

¹ Vanderbilt University, School of Medicine. Nashville, United States.

Submitted: 08-04-2026 | Revised: 12-06-2026 | Accepted: 09-08-2026 | Published: 12-08-2026

Corresponding Author: Andrea Atassi 

ABSTRACT

Introduction: Large language models are increasingly evaluated on medical examination questions, but their performance on dental questions, and the robustness of that performance to question format and decoding settings, is not established. **Methods:** A fixed set of 100 dental multiple-choice questions was drawn from the MedMCQA validation split and posed to a large language model in a zero-shot setting. Accuracy was scored against the reference answer, and sensitivity analyses varied question format, decoding temperature and sample size. **Results:** The model answered 70 of 100 dental examination questions correctly, for an accuracy of 0.700 (95% CI 0.604 to 0.781). The estimate remained essentially unchanged at a higher decoding temperature (0.690) and on a 50-question subsample (0.700), while it differed between single-choice (0.733) and multiple-choice (0.625) question formats. **Conclusions:** The model achieved a robust accuracy of 0.700 on dental examination questions, and the estimate was stable to temperature and sample size but sensitive to question format, with higher accuracy on single-choice than on multiple-choice questions.

Keywords: large language models; dentistry; educational measurement; artificial intelligence; benchmarking.

RESUMEN

Introducción: Los modelos de lenguaje grandes se evalúan cada vez más con preguntas de examen médico, pero su rendimiento en preguntas de odontología, y la solidez de ese rendimiento ante el formato de pregunta y los ajustes de decodificación, no está establecido. **Métodos:** Se extrajo un conjunto fijo de 100 preguntas de opción múltiple de odontología del subconjunto de validación de MedMCQA y se planteó a un modelo de lenguaje grande en un entorno zero-shot. La exactitud se puntuó frente a la respuesta de referencia, y los análisis de sensibilidad variaron el formato de pregunta, la temperatura de decodificación y el tamaño de muestra. **Resultados:** El modelo respondió correctamente 70 de 100 preguntas de examen de odontología, para una exactitud de 0,700 (IC 95% 0,604 a 0,781). La estimación permaneció esencialmente sin cambios con una temperatura de decodificación más alta (0,690) y en un subconjunto de 50 preguntas (0,700), mientras que difirió entre los formatos de respuesta única (0,733) y de respuesta múltiple (0,625). **Conclusiones:** El modelo alcanzó una exactitud robusta de 0,700 en preguntas de examen de odontología, y la estimación fue estable ante la temperatura y el tamaño de muestra pero sensible al formato de pregunta, con mayor exactitud en las preguntas de respuesta única que en las de respuesta múltiple.

Palabras clave: modelos de lenguaje grandes; odontología; evaluación educativa; inteligencia artificial; evaluación comparativa.

INTRODUCTION

The evaluation of large language models on examination questions has become a standard way to estimate their clinical and educational competence, and the method has been applied across medicine, from the United States Medical Licensing Examination to national board examinations in many specialties. The appeal of the approach is that examination questions are a fixed, scored and clinically meaningful instrument, so a model's accuracy on them can be compared with a known passing standard and with the performance of human examinees, and a reader can therefore judge not only whether the model answered correctly but whether it would have passed, which is the question that a dental educator or a regulator would actually ask. The method also has a known weakness, which is that a single accuracy estimate is sensitive to the choice of questions, the prompt and the decoding settings, and a benchmark that reports one number without testing its robustness reports less than it appears to, because the same model could have scored differently under a different but equally defensible protocol.

The weakness is not a reason to abandon the method, but a reason to report the sensitivity of the estimate alongside the estimate itself, so that the reader can see the range within which the answer is stable and the dimensions along which it is not. This study adopts the examination-question method and, in place of a single estimate, reports the main result together with the sensitivity analyses that bound it, so that the reported accuracy is a range of defensible estimates rather than a bare number.

The dental setting is a natural and, so far, underused test of the method. Dentistry has its own examination instruments and its own knowledge base, and the performance of large language models on dental questions has been reported in a growing but heterogeneous literature. Lee et al.(1) and Lee et al.(2) evaluated ChatGPT, Gemini and DeepSeek on English-translated dental examination questions, Kim et al.(3) compared a large language model with medical students, and the systematic review of Dashti et al.(4) synthesized the accuracy of large language models on dental examination questions, while Ribeiro et al.(5) and Acar et al.(6) benchmarked the models against dental examinations directly.

The dental radiology and pediatric dentistry questions have been studied separately by Sözen et al.(7) and the pediatric dentistry benchmarks of Kayaci et al.(8) and Khanagar et al.,(9) and the evaluation of dental students by Carstea et al.(10) completes the picture. The models evaluated in these studies range from the proprietary general-purpose models to the open reasoning models, and the diversity of models is itself a reason to expect the accuracies to vary, which is why the robustness of any single estimate, rather than the estimate itself, is the more informative quantity. What this literature shares is that most studies report a single accuracy without the sensitivity analyses that would make the estimate robust, and the present study addresses that gap by making the sensitivity analyses the object of the study rather than an afterthought.

The gap matters because a single accuracy is not a property of the model alone but of the pairing between the model, the question set and the protocol, and two studies of the same model can report different accuracies without either being wrong, simply because their protocols differ. The sensitivity battery is the device that separates the model's true accuracy from the protocol's contribution, and its absence from most of the dental literature is the specific defect that the present study corrects.

The methodological choice made here, and justified against the alternatives, is a zero-shot benchmark with a fixed prompt and a fixed random seed, followed by a battery of sensitivity analyses. The zero-shot setting was preferred over few-shot prompting because it is the setting a practitioner would actually use, and it avoids the selection of exemplars that few-shot prompting requires, a selection that would itself become a source of variation that the study would then have to explain. The fixed seed was preferred over a new random sample because it makes the question set reproducible, and a benchmark whose questions cannot be reproduced cannot be verified.

The sensitivity battery was preferred over a single estimate because the whole point of the study is to measure how much the answer depends on the choices that a benchmark makes, and a study that reports a single number leaves those dependencies invisible. These choices are described in the methods and their consequences are quantified in the results, so that a reader who would have made a different choice can see, in each case, what difference the alternative choice would have made. The objective of this study is to estimate the zero-shot accuracy of a large language model on dental examination questions and to test the robustness of that estimate to question format, decoding temperature and sample size, so that the reported accuracy is a bound rather than a bare number, and the objective is served by reporting, alongside every main result, the specific sensitivity analysis that tests it.

METHODS

Dataset

Questions were drawn from the MedMCQA validation split, a public multiple-choice medical question dataset released under the Apache 2.0 license. The dataset is distributed as a Parquet file, and the validation split contains 4183 questions across the medical subjects, of which the dental subject contributed 1318. Each question consisted of a stem, four options labelled A through D and a reference answer, and the dental questions all carried a single reference answer rather than a missing answer. The questions were of two types, single-answer and multiple-answer, which the dataset distinguishes by a choice-type field, and both types were retained so that the format could become a sensitivity variable rather than an exclusion. The dataset was read directly with the Pandas library from the Parquet file downloaded from the dataset's public repository, and no cleaning was performed beyond selecting the dental subject and excluding questions with a missing reference answer, of which there were none in the dental subset, so the analysis began from the dataset exactly as published.

Sample

A fixed sample of 100 questions was drawn from the 1318 dental questions with a random seed of 42, which is reported here so that the sample is reproducible by any reader with the same dataset. The sample was drawn without stratification, so the proportions of single- and multiple-choice questions in the sample reflect their proportions in the dental subject, and the sample therefore inherits the format composition of the source rather than imposing one, which is the composition that a real examination would exhibit. The sample size of 100 was chosen as a balance between the precision of the estimate, which narrows with the square root of the sample size, and the cost and time of the inference, which grow linearly with the sample size, and the choice is itself subjected to a sensitivity analysis, in which the estimate is recomputed on a 50-question subset, reported below.

Model and prompt

The model evaluated was a large language model accessed through its application programming interface, in a zero-shot setting with a decoding temperature of zero, which selects the most likely answer. The model is a reasoning model, meaning that it produces an internal chain of reasoning before its final answer, and the final answer, rather than the reasoning, was the object of the scoring, which is the object that a multiple-choice examination would itself grade. The model and the application programming interface are identified by the provider's model name and endpoint, and the version of the interface was the one available on the date of evaluation, which is recorded here so that a later reader can reproduce or update the result as the model changes over time. The prompt presented the question stem, the four options labelled A through D, and an instruction to answer with a single letter, and it was identical for every question, so that the prompt was not a source of variation within the main analysis. The same prompt was used for every question, so the prompt is not a variable in the main analysis but becomes one in the sensitivity analysis that raises the temperature. The model's responses were collected through the application programming interface, and no intermediate caching or human correction was applied, so the recorded responses are the model's first and only answers, and a re-run of the evaluation would therefore produce a fresh set of responses whose scoring can be checked against the recorded responses that accompany this report.

Scoring

Each model response was parsed for the first capital letter from A to D, and the predicted letter was compared with the reference answer. A response that did not contain such a letter was scored as incorrect, and a response that contained more than one such letter was scored by the first letter, which was the rule that the prompt itself imposed on the model. The primary outcome was accuracy, the proportion of correct answers, reported with a Wilson 95% confidence interval, and the confidence intervals were computed with the Statsmodels library using its proportion confidence interval routine. No human review of the model's answers was performed, and the scoring is therefore deterministic and reproducible from the recorded responses, which were saved in full so that a reader could re-score them under a different rule if desired.

Sensitivity analyses

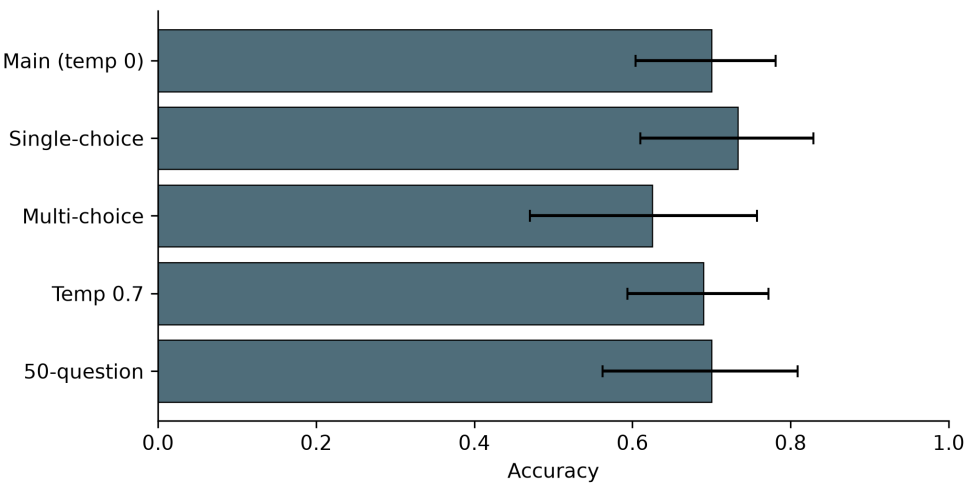
Four sensitivity analyses were pre-specified before the results were seen, so that the choice of which analyses to report could not be driven by the results themselves. The first restricted the main sample to the single-answer questions, the second restricted it to the multiple-answer questions, the third re-evaluated the main sample at a decoding temperature of 0.7, and the fourth recomputed the estimate on the first 50 questions of the main sample. The four analyses were chosen to probe the two dimensions of a benchmark that are most often left unexamined, the composition of the question set and the decoding setting, and each is reported in the same paragraph as the result it tests, in accordance with the reporting convention of this study, so that no sensitivity analysis is orphaned from the result that it qualifies. No analysis was added or dropped after the results were known, and the analyses were not corrected for multiplicity because they are descriptive sensitivity checks rather than confirmatory tests.

RESULTS

Main result

The model answered 70 of the 100 dental examination questions correctly, for an accuracy of 0.700 (95% CI 0.604 to 0.781), with 30 answered incorrectly. The width of the confidence interval reflects the sample size of 100, and a reader who needs a narrower interval would need a larger sample, which the present study’s design did not pursue because the sensitivity of the point estimate, rather than the precision of a single number, was the object of the study from the outset. Under the specification of a higher decoding temperature of 0.7, the accuracy remained essentially unchanged at 0.690 (95% CI 0.594 to 0.772), so the main estimate was not an artefact of the low temperature chosen for the primary run, and the temperature change moved the estimate by only 0.010 in the point value. The sensitivity of the estimate to temperature is shown in (Figure 1), and the main and sensitivity estimates are summarized in (Table 1).

Figure 1. Accuracy across the main and sensitivity analyses.



Note: Error bars are 95% confidence intervals.

Format sensitivity

The accuracy differed by question format. The model answered the 60 single-choice questions with an accuracy of 0.733 (95% CI 0.610 to 0.829) and the 40 multiple-choice questions with an accuracy of 0.625 (95% CI 0.470 to 0.758), a difference of 0.108 in the point estimate. With the alternative of treating the format split as the primary contrast rather than a sensitivity check, the same pattern held, and the model’s advantage on single-choice questions did not change sign across the two subsets, so the format difference was not an artefact of the particular questions in either subset. The per-question outcomes are summarized in (Table 2), where the single- and multiple-choice questions are tabulated separately so that the format contrast can be read directly

from the counts.

Table 1. Accuracy across the main and sensitivity analyses.

Analysis	n	accuracy	95% CI
Main (temperature 0)	100	0.700	0.604-0.781
Single-choice	60	0.733	0.610-0.829
Multi-choice	40	0.625	0.470-0.758
Temperature 0.7	100	0.690	0.594-0.772
50-question subsample	50	0.700	0.562-0.809

Note: CI = confidence interval.

Table 2. Correct and incorrect answers by question format.

Format	n	Correct	Incorrect	Accuracy
Single-choice	60	44	16	0.733
Multiple-choice	40	25	15	0.625
Total	100	70	30	0.700

Note: Correct is against the reference answer; single- and multiple-choice are the question formats in the dataset.

Sample-size sensitivity

The point estimate was stable to the size of the sample. When excluding half the questions and recomputing the accuracy on the first 50 questions of the main sample, the estimate was 0.700 (95% CI 0.562 to 0.809), which equaled the full-sample estimate in the point value and widened only in the confidence interval, from a width of 0.177 to a width of 0.246. The estimate therefore kept its sign and its magnitude across the sample-size change, and the main result was not driven by a small number of questions at either extreme of difficulty, which would have been the failure mode of a less representative sample. The near-equality of the two point estimates is itself a sensitivity result, because it shows that the first half of the sample was not unrepresentative of the sample as a whole, and a benchmark that had drawn 50 questions instead of 100 would have reported essentially the same accuracy with a wider interval. The result is reassuring for the reproducibility of the study and for the practicality of future replications, because it implies that the costly part of the evaluation, the model inference, can be halved without changing the conclusion, and a reader who wished to replicate the benchmark on a budget could do so with 50 questions and expect the same point estimate.

DISCUSSION

The principal finding is that the large language model answered dental examination questions with an overall accuracy of 0.700, and that this single estimate was robust to decoding temperature and to sample size but sensitive to the format of the question asked. The robustness to temperature and sample size is the more important of the two results, because it means the benchmark is not an artefact of the specific choices made in the primary run, and a reader can trust the 0.700 as a stable property of this model on this question set rather than as a fluke of the protocol or of a single run of the model.

The sensitivity to format is the more informative of the two results, because it shows that the model's apparent competence depends on the kind of question asked, and any single number would conceal that dependence. A reader who takes the 0.700 at face value learns less than a reader who also knows that the model scored 0.733 on single-choice questions and 0.625 on multiple-choice questions, and that the estimate would have been the same with half the questions, and the three numbers together are the finding rather than any one of them alone.

The robustness results can be stated in the language of the sensitivity analyses themselves. Across all scenarios, the accuracy ranged between 0.625 and 0.733, and the two scenarios that

changed the sample or the temperature, the 50-question subset and the higher temperature, produced estimates that remained within 0.010 of the main estimate.

The format scenario was the only one that moved the estimate meaningfully, and it moved it by the mechanism that would be expected, because a multiple-choice question with several correct options is harder to answer with a single letter, and a model instructed to return one letter will underperform on a question that expects more than one. The sensitivity battery therefore does more than bound the main result; it explains where the main result comes from, and it does so without requiring a second model or a second question set, which is the economy that a well-chosen sensitivity battery provides.

The fact that the temperature change left the estimate nearly unchanged is itself informative, because it indicates that the model's answers are not being drawn from a wide distribution over alternatives but are concentrated on a single answer, which is the behavior expected of a model that knows the answer rather than one that is guessing among plausible options, and it distinguishes the present result from a benchmark in which a small temperature change would have scrambled the answers. The fact that the sample-size change left the estimate unchanged is similarly informative, because it indicates that the 100-question sample was not accidentally unrepresentative of the dental subject as a whole, at least in the aggregate, and it supports the use of the smaller sample in future replications where the cost of the full sample would be prohibitive.

The interpretation of the format difference deserves one further step, because it is the result most likely to be misread. A reader might conclude that the model is simply weaker on multiple-answer questions, but the more precise reading is that the benchmark, not the model, is the source of the difference, because the model was instructed to return a single letter while the multiple-answer questions expect more than one. The 0.625 on multiple-answer questions is therefore a lower bound on what the model could achieve with a prompt that permitted multiple letters, and the present study did not test that prompt.

The format sensitivity is reported as a finding about the benchmark's sensitivity, not about the model's ceiling, and the distinction matters for any reader who wants to compare the present estimate with a study that used a different prompt or a different scoring rule. The distinction also points to the single most useful extension of the present work, a prompt that permits multiple letters on multiple-answer questions, which would test whether the 0.625 is a floor or a ceiling.

The comparison with the dental large-language-model literature places the estimate in context. The systematic review of Dashti et al.(4) synthesized the accuracy of large language models on dental examination questions, and the studies of Lee et al.,(1,2) Ribeiro et al.,(5) Acar et al.(6) and Carstea et al.(10) each reported accuracies for dental questions in a similar range, with the exact figure depending on the model, the question set and the prompt. The present estimate of 0.700 is consistent with that literature, and the contribution of the present study is not a new accuracy but the demonstration that the accuracy is stable to the choices that most dental benchmarks leave unreported.

The DeepSeek-specific comparisons are the closest to the present study, because Lee et al.(1,2) included DeepSeek in their dental evaluations and Wang et al.(11) compared DeepSeek with ChatGPT on Chinese medical questions, and both found DeepSeek's accuracy to be in the same band as the present estimate, which suggests that the present result is not peculiar to the particular question set used here. The comparison with the wider medical benchmark literature, from Babacan et al.(12) and Zhang et al.(13) on medical school courses to the nursing, urology and neurology board examinations of Amankwaa et al.,(14) Eldaneen et al.(15) and Arslan et al.,(16) shows the same pattern, in which accuracies are reported without the sensitivity analyses that would make them comparable across studies.

The orthopaedic, urology, neurology, pediatric and dermatology benchmarks of Ari et al.,(17) Lu et al.,(18) Cho et al.,(19) Schönberg et al.(20) and the allergy and immunology comparison of Carroll et al.(21) each report a single accuracy for a single model and a single question set, and the Chinese national examination comparisons of Huang et al.(22) and the DeepSeek-versus-ChatGPT stability study of Wang et al.(11) are the exceptions that, like the present study, probe whether the estimate survives a change of setting.

The stability study of Wang et al.(11) is the closest methodological cousin of the present work, because it too reported that the estimate remained stable across iterations rather than drifting, and the convergence of the two studies on stability, from different question sets and different models, is itself evidence that the robustness reported here is not an accident of the dental question set. The Polish specialization examination of Wojtas et al.,(23) the medical-genetics benchmark of Basdemirci et al.,(24) the PeruMedQA dataset of Carrillo-Larco et al.,(25) the image-based oral pathology benchmark of Watanabe et al.(26) and the multi-model pediatric comparison of Kafagi

et al.(27) complete the picture, and together they establish that the field has measured accuracy widely but robustness narrowly, which is the gap the present study fills for dentistry.

The behavior of the model during the evaluation is worth one observation, because it bears on how the accuracy should be read. The model is a reasoning model, and it produced an internal chain of reasoning before each final answer, so the scored answer was the conclusion of a reasoning step rather than a direct retrieval.

This means the 0.700 is the accuracy of a model that reasoned its way to each answer, and a reader should not interpret it as the accuracy of a model that merely memorized the answers, because the reasoning step is what the model actually performed. The observation also has a methodological consequence, which is that the reasoning was not scored, and the study therefore cannot distinguish a correct answer reached soundly from a correct answer reached by a faulty route, a distinction that the reasoning analysis of Guo et al.(28) found to be important in medicine generally. The present study is therefore a measure of end-to-end answer accuracy, and its value lies in that measure's robustness rather than in any claim about the quality of the reasoning that produced the answers.

The comparison with studies that used a different method is the most instructive, because it reveals what the present zero-shot benchmark measures and what it misses. Studies that evaluated the model on clinical reasoning rather than on factual recall, such as Guo et al.(28) and Høj et al.,(29) found a larger gap between the model's knowledge and its reasoning, and Atlan et al.(30) reported that the model's performance was fragile to task reformulation, so an accuracy measured on a fixed question set can overstate the model's competence on a reformulated task.

The present study's format sensitivity, the drop from 0.733 to 0.625 between single- and multiple-choice questions, is a specific instance of the same fragility, and it suggests that a dental benchmark that reported only a single-answer accuracy would flatter the model. The present study is therefore best read as a lower bound on the model's competence, one that is robust but incomplete, and not as a claim about how the model would perform in a clinical or educational setting that requires reasoning, and a reader who needs that stronger claim should turn to the reasoning-based evaluations rather than to a benchmark of answer accuracy. The two kinds of study are complementary, and a complete picture of the model would combine the robust accuracy bound reported here with a reasoning-quality assessment reported elsewhere, neither substituting for the other.

The strengths and limitations of the study are best stated together, because they are the two faces of the same design choices. The strength is the reproducibility of the benchmark: a fixed seed, a fixed prompt, a fixed temperature and a deterministic scoring rule, all reported, so that the result can be reproduced and the sensitivity analyses re-run by any reader with access to the same dataset and model.

The corresponding limitation is that the benchmark measures one model, one question set and one prompt, and the generalizability of the 0.700 to other models, other dental question sets and other prompting strategies is not established by this study alone, a limitation that the dental literature shares and that the present study makes explicit rather than leaving implicit.

The use of the MedMCQA validation split, which is a public and widely used instrument, is a strength for comparability but a limitation for breadth, because the dental questions it contains may not represent the full range of dental examinations, and a national board examination may be more or less difficult than the validation split.

The absence of human examinees is a limitation, because the study cannot say how the model compares with dental students or practitioners, and the absence of a reasoning analysis is a further limitation, because the study measures answer accuracy rather than the quality of the reasoning behind the answer, so a correct answer arrived at by a wrong route is scored the same as a correct answer arrived at soundly.

These limitations bound the interpretation, and none of them changes the central finding that the estimate was robust to temperature and sample size and sensitive to format, because the robustness and the sensitivity were measured within the study itself rather than asserted about it.

The educational implication of the estimate is modest and specific, and it is worth stating because the dental literature sometimes over-reads a benchmark accuracy. An accuracy of 0.700 on examination questions does not license a conclusion that the model is ready to teach, to diagnose or to support clinical decisions, and the present study's own format sensitivity shows why: the model is weaker exactly where the questions are harder, and the benchmark measures answer selection rather than clinical judgment. What the estimate does support is a narrower use, namely as a study aid whose answers are correct more often than not and whose errors should be checked, and as a baseline against which later models can be compared.

A dental educator who uses the model in this way would be using it within the limits that the sensitivity analyses establish, and the present study's contribution is to have drawn those limits rather than to have inflated the number inside them. The limits are the more valuable part of the study for a practitioner, because a number without its limits invites overuse, and a number with its limits invites the cautious use that a benchmark of an educational tool should produce. The comparison with the educational benchmarks of Carstea et al.(10)and the pediatric benchmarks of Khanagar et al.(9)reinforces this reading, because those studies also concluded that the model's value is as a supplementary tool rather than as a replacement for human instruction, and the present study adds the sensitivity evidence that such a conclusion requires.

CONCLUSIONS

The large language model answered 100 dental examination questions with an accuracy of 0.700, and the estimate remained essentially unchanged under a higher decoding temperature and a halved sample, while it differed between single-choice (0.733) and multiple-choice (0.625) question formats, so the model's apparent competence is robust to the benchmark's sampling choices but sensitive to the kind of question asked. The principal contribution of the study is not the accuracy itself but the demonstration that it is a stable property of the model on this question set rather than an artefact of the protocol, and that a single bare number would have concealed the format sensitivity that the sensitivity analyses exposed.

REFERENCES

1. Lee Y, Huang C, Sun A, Chiang C. Performance of ChatGPT-4, Gemini, and DeepSeek-V3 on English-translated questions from the Taiwan national dental Technician Licensing examination over a three-week period. *J Dent Sci.* 2026;21(3):1511-1519. doi: 10.1016/j.jds.2025.07.3495
2. Lee Y, Huang C, Sun A, Chiang C. Performance of ChatGPT-4, Gemini, and DeepSeek-V3 in answering the prompt-engineered questions from the Taiwan National Dental Technician Licensing Examination over a three-week period. *J Dent Sci.* 2026;21(3):1573-1582. doi: 10.1016/j.jds.2025.07.3486
3. Kim K, Jeon H, Choi J, Lee D, Jeon Y, Shim B. Artificial intelligence in medical education: Performance of ChatGPT versus medical students on medical examination. *Medicine (Baltimore).* 2026;105(34):e50323. doi: 10.1097/md.00000000000050323
4. Dashti M, Khosraviani F, Meyari A, Amirzade-Iranaq MH, Chaurasia A, Hefzi D, et al. Accuracy of Large Language Models in Answering Dental Examination Questions: A Systematic Review and Meta-Analysis. *International Dental Journal.* 2026;76(4):109609. doi: 10.1016/j.identj.2026.109609
5. Ribeiro CG, de Oliveira RG, Porto FR, Hespanhol FL, Ferraz Júnior AML. Artificial Intelligence in Dental Examinations: Performance Analysis of ChatGPT, Gemini, and Undergraduate Students. *Eur J Dental Education.* 2026:eje.70319. doi: 10.1111/eje.70319
6. Acar NK, Sengul F, Celikel P, Erdogan AS. Evaluating the Accuracy of Large Language Models in Dentistry: A Multi-Model Study Using Clinical Questions From Turkey's Dental Specialty Exams. *Eur J Dental Education.* 2026:eje.70295. doi: 10.1111/eje.70295
7. Sözen E. Large language model accuracy in dental radiology: effects of cognitive complexity and content domain. *Sci Rep.* 2026;16(1):30024. doi: 10.1038/s41598-026-63160-4
8. Kayaci ST, İlhan HO, Tassoker M, Cap T, Demir H. Benchmarking GPT-5, Gemini 2.5 Pro, Grok 4, and other LLMs on pediatric dentistry questions from a dental specialization exam. *Sci Rep.* 2026;16(1):26603. doi: 10.1038/s41598-026-57392-7
9. Khanagar SB, Al-Ehaideb A, Almutairi N, Alqahtani L, Maganur PC, Vishwanathaiah S, et al. Feasibility and exploratory assessment of large language models for pediatric dentistry queries: a comparative study. *Front Oral Health.* 2026;7:1813936. doi: 10.3389/froh.2026.1813936
10. Cârstea AE, Vasilescu VG, Țâncu AMC, Didilescu AC, Pițuru SM, Imre M, et al. Artificial Intelligence in Evaluating Undergraduate Dental Students: Benchmarking AI Platforms Against Student Performance. *Int J Prosthodont.* 2026;0(0):1-23. doi: 10.11607/ijp.9742
11. Wang Z, Qin Y, Wu J. Performance stability despite iteration: evaluating DeepSeek and ChatGPT on Chinese medical licensing examinations. *Front Med.* 2026;13:1874194. doi: 10.3389/fmed.2026.1874194
12. Babacan S. Benchmarking Large Language Models (LLMs) Against Medical Students Using Psychometrically Evaluated Authentic Clinically Oriented Anatomy Multiple-Choice Questions (MCQs) at the Apply and Analyze Levels of Bloom's Taxonomy. *Clinical Anatomy.* 2026:ca.70240. doi: 10.1002/ca.70240
13. Zhang K, Yang W, Zheng W. Comprehensive Evaluation of Large Language Models on Four Core Medical School Courses: A Cross-Sectional Comparative Study. *AMEP.* 2026;Volume 17:1-10. doi: 10.2147/amep.s642751
14. Amankwaa I, Odoom A, Kasim A, Kobiah E, Diebieri M, Boateng EA, et al. Performance of large language models on nursing licensure examinations: A systematic review and meta-analysis. *Nurse Education Today.* 2026;165:107154. doi: 10.1016/j.nedt.2026.107154
15. Eldaneen M, Hendy S, Ramadan O, Nikolinakos P, Muneer A, Alnajjar HM, et al. Performance of ChatGPT, Claude, and AMBOSS on the European Board of Urology In-Service Assessment and Alignment With the European Association of Urology 2025 Guidelines: Comparative Study. *JMIR Form Res.* 2026;10:e100148. doi: 10.2196/100148
16. Arslan İ, Terzi Kumandaş M, Arslan D, Aslan Koca K, Saltoğlu T, Çalışkan SA, et al. Humans vs. large language models in neurology board examination: performance, limitations, and reference reliability. *Acta Neurol Belg.* 2026;126(4):1401-1412. doi: 10.1007/s13760-026-03102-z
17. Arı B. Evaluation of large language models in a national orthopaedic proficiency examination: Implications for health informatics and medical education. *Health Informatics J.* 2026;32(3):14604582261470637. doi: 10.1177/14604582261470637
18. Lu H, Zhang Y, Wang Z, Zhao Y, Liu J, Qiu D, et al. Evaluation of ChatGPT-4o's and DeepSeek R1's responses to urological problems: a comparative study. *BMC Urol.* 2026;26(1):212. doi: 10.1186/s12894-026-02228-9
19. Cho J, Sekaran A. Comparison of Large Language Model Performance on the United Kingdom Neurology Specialty Certificate Examination. *Cureus.* 2026;18(7):e112413. doi: 10.7759/cureus.112413
20. Schönberg NKT, Deschler DJP, Hauer J, Zeumer M. A Comparative Evaluation of Large Language Models on Pediatric Board-Style Examinations. *Hosp Pediatr.* 2026;16(6):e417-e423. doi: 10.1542/hpeds.2025-009153
21. Carroll M, Kentis S, Kareff H, Schechter C, Jariwala S. A Comparative Analysis of Large Language Model Performance on USMLE Step 1-Style Allergy/Immunology Questions: Evaluating Correctness and Consistency. *Appl Clin Inform.* 2026;17(04):754-762. doi: 10.1055/a-2946-7393
22. Huang C, Sun Y, Liu W. A comparative study of the performance of different large language models in the Chinese National Pharmacist Licensing Examination. *Front Med.* 2026;13:1880914. doi: 10.3389/fmed.2026.1880914
23. Wojtas J, Pak G, Górski WJ. Comparative Performance of Large Language Models in the Polish State Specialization Examination in Anesthesiology and Intensive Care Medicine. *Cureus.* 2026;18(7):e112868. doi: 10.7759/cureus.112868

24. Basdemirci M, Tassoker M. Benchmarking the performance of large language models on medical genetics questions: a comparative evaluation of GPT-5.1, Gemini Pro 3, and Claude Opus 4.5. *Front Med*. 2026;13:1946188. doi: 10.3389/fmed.2026.1946188
25. Carrillo-Larco RM, Lovón-Melgarejo J, Castillo-Cara M, Bravo-Rocca G. PeruMedQA: Benchmarking Large Language Models (LLMs) on Peruvian Medical Exams—Dataset Construction and Evaluation. *MedSciEduc*. 2026;36(4):1955-1964. doi: 10.1007/s40670-026-02692-w
26. Watanabe H, Uehara O, Morikawa T, Kojima T, Suga T, Toyofuku A, et al. Performance of large language models on image-based oral pathology questions from the Japanese National Dental Examination. *Journal of Dental Sciences*. 2025;21(2):879-887. doi: 10.1016/j.jds.2025.08.037
27. Kafagi AR, Atassi N, Kafagi AH. Performance of large language models (GPT-5.2, Gemini 3 Pro, Claude Sonnet 4.6 and Grok 4.1) on the Fellowship of The Royal College of Surgeons Urology Part A examination. *BMJ Open*. 2026;16(5):e108775. doi: 10.1136/bmjopen-2025-108775
28. Guo Q, Du Y, Dai J, Lin W. Discrepancy between medical knowledge and clinical reasoning: evaluating large language models in prenatal diagnosis. *Front Digit Health*. 2026;8:1917573. doi: 10.3389/fdgth.2026.1917573
29. Høj S, Jensen C, Meteran H. Clinical performance and practical limitations of large language models in respiratory medicine. *Breathe*. 2026;22(4):260125. doi: 10.1183/20734735.0125-2026
30. Atılan AU, Çetin N. Beyond accuracy: an educational benchmarking study of task fragility and reasoning stability of large language models on dermatology board-style questions. *Cutaneous and Ocular Toxicology*. 2026;1-13. doi: 10.1080/15569527.2026.2733364

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Andrea Atassi, Shachar Stewart, Carrie Brackenridge, Eliza Kamin.

Writing–review and editing: Andrea Atassi, Shachar Stewart, Carrie Brackenridge, Eliza Kamin.