



Disparidades lingüísticas en el rendimiento de un modelo de lenguaje grande en preguntas dentales en siete idiomas

Language disparities in large language model performance on dental questions across seven languages

Maikro Osvaldo Chávez Moya¹  

Leodanis Hernández Cabrera¹  

Neisy Pérez Ramos²  

¹ Universidad de Ciencias Médicas de Villa Clara, Facultad de Medicina. Santa Clara, Cuba.

² Universidad de Ciencias Médicas de Villa Clara, Facultad de Medicina, Departamento de Ciencias Biomédicas. Santa Clara, Cuba.

Submitted: 13-06-2026 | Revised: 21-08-2026 | Accepted: 14-10-2026 | Published: 18-10-2026

Corresponding Author: Maikro Osvaldo Chávez Moya 

ABSTRACT

Introduction: Large language models answer dental questions, but their accuracy across languages has not been measured. **Methods:** Sixty dental multiple-choice questions from MedMCQA were translated into Swahili, Hindi, Arabic, Chinese, Portuguese and Spanish with a fixed protocol and answered by the DeepSeek model at temperature zero, with English as the reference. **Results:** Accuracy was 61.7% in English and 61.7% in Portuguese, and it fell to 60.0% in Chinese, 58.3% in Arabic and Spanish, 55.0% in Hindi, and 26.7% in Swahili. The English-to-language gap reached 6.7 percentage points for Hindi and 35.0 points for Swahili. **Conclusions:** Accuracy held near the English reference in the high-resource European and Asian languages but collapsed in Swahili, so the language disparity in dental question answering is concentrated in the low-resource language rather than spread evenly across languages.

Keywords: large language model; language disparity; dental question answering; machine translation; multilingual evaluation.

RESUMEN

Introducción: Los modelos de lenguaje grandes responden preguntas dentales, pero su exactitud entre idiomas no se ha medido. **Métodos:** Sesenta preguntas dentales de opción múltiple de MedMCQA se tradujeron al suajili, hindi, árabe, chino, portugués y español con un protocolo fijo y fueron respondidas por el modelo DeepSeek a temperatura cero, con el inglés como referencia. **Resultados:** La exactitud fue del 61,7 % en inglés y del 61,7 % en portugués, y cayó al 60,0 % en chino, 58,3 % en árabe y español, 55,0 % en hindi y 26,7 % en suajili. La brecha respecto al inglés alcanzó 6,7 puntos porcentuales en hindi y 35,0 puntos en suajili. **Conclusiones:** La exactitud se mantuvo cerca de la referencia en inglés en los idiomas europeos y asiáticos de altos recursos, pero se desplomó en suajili, de modo que la disparidad lingüística en la respuesta a preguntas dentales se concentra en el idioma de bajos recursos y no se reparte por igual entre idiomas.

Palabras clave: modelo de lenguaje grande; disparidad lingüística; respuesta a preguntas dentales; traducción automática; evaluación multilingüe.

INTRODUCTION

Large language models have moved from general benchmarks to the evaluation of clinical knowledge, and their performance on medical question answering is now documented on standardized examinations and clinical corpora. The trajectory began with broad knowledge benchmarks such as the Massive Multitask Language Understanding suite, which established that models could answer questions across many disciplines, and it reached medicine when Singhal et al.(1) demonstrated that a model tuned for clinical knowledge could approach expert-level answers.(2)

The subsequent benchmarking of general-purpose models on the United States Medical Licensing Examination established that commercial chatbots perform near or above the passing threshold, and the evaluation of stronger models on medical challenge problems extended that result to a wider set of tasks.(3-5)

These results collectively created the expectation that a language model could answer a patient's or a student's medical question directly, and that expectation has extended to dentistry, where models are now compared on dental implantology inquiries, dental licensing examinations and oral-anatomy education.(6-8)

The biomedical applications of these models have in turn become a field of their own, surveyed both for their general capabilities and for the specific challenge of medical question answering, and the surveys consistently note that the models are evaluated and deployed in English far more than in any other language.(9,10)

That asymmetry is the precise starting point for the present study: the evidence for a model's dental competence is overwhelmingly evidence of its English competence, and the question of whether that competence travels to other languages is left entirely open by the existing literature.

That expectation, however, was formed almost entirely in English. The corpora on which the strongest models are trained are dominated by English text, and the multilingual behaviour of a model is known to lag its English behaviour, a gap that surveys of multilingual models attribute to unequal training data across languages, with the languages of the Global South and of Africa especially underrepresented.(11,12)

Bang et al.(13) showed that ChatGPT's performance varied across languages on a broad set of tasks, and the multilingual medical literature has begun to document the same phenomenon, with language-dependent accuracy reported for diagnostic questions and for board examinations in Turkish and other languages.(14,15) The biases that accompany this disparity have been catalogued in multilingual clinical corpora, where a model's behaviour shifts across languages in ways that are not reducible to translation quality alone, and where the same clinical scenario can receive different treatment depending on the language in which it is posed.(16)

The concern is not merely academic: a model that answers correctly in English but errs in Swahili or Hindi would widen, rather than narrow, the very disparities in access to health information that language technology is often invoked to reduce, and the population most likely to rely on a free language-model answer is the population least likely to have that answer be correct.(17)

Dental question answering is a case where that concern is concrete and, until now, largely unstudied. Dental applications of language models have proliferated, and the dental literature now includes evaluations of general models on implantology, Chinese licensing examinations and multilingual terminological support, but these studies measure English or a single additional language rather than a systematic cross-linguistic comparison.(18-20)

The choice of languages in that literature reflects the same English-centrism as the broader field, with the consequence that a dental model's behaviour in the low-resource languages spoken across large parts of Africa, Asia and Latin America is essentially undocumented, and a model is routinely described as competent on dental tasks on the strength of evidence gathered in a single language.

From English to the six other languages examined here, the accuracy of a dental question-answering model has not been traced, and the size of the accuracy gap between English and each translated language remains unknown. This study addressed that gap by evaluating a large language model on a fixed sample of dental multiple-choice questions in English and in translations into Swahili, Hindi, Arabic, Chinese, Portuguese and Spanish, so that the accuracy trajectory from the reference language to the translated languages could be measured directly. The objective was to measure the model's accuracy in English and in the six translated languages and to quantify the accuracy gap between English and each translated language.

METHODS

Dataset and sample

The evaluation used the Dental subset of the MedMCQA dataset, a large multi-subject multiple-choice dataset for medical question answering that has served as a standard benchmark for the medical-domain evaluation of language models.⁽²¹⁾ The Dental subset contained 1,318 multiple-choice questions drawn from dental examinations, each with four options and a single gold answer, and the questions spanned single-answer and multi-answer formats, which allowed the evaluation to cover the two formats used in dental examinations.

A fixed random sample of 60 dental questions was drawn for the evaluation, with the sample chosen to be small enough to keep the translation and evaluation calls tractable while large enough to produce a stable per-language accuracy estimate, and the sample was stratified by answer format so that it retained the single-answer to multi-answer ratio of the full subset. The same 60 questions were used in every language, so that the comparison between languages was paired by question and not confounded by differences in the sampled questions, and any difference in accuracy between two languages reflects the language of presentation alone rather than a difference in question difficulty or in the set of questions posed.

Translation protocol

Each question and its four options were translated from English into six languages, namely Swahili, Hindi, Arabic, Simplified Chinese, Portuguese and Spanish, chosen to span the African, Asian, European and Latin American language families that the study sought to compare and to include at least one genuinely low-resource language, Swahili, alongside the high-resource languages that dominate existing multilingual evaluations.

The translation was performed by the same model under evaluation, using a fixed prompt that instructed the model to translate the question and its options without answering them, to preserve the option labels A, B, C and D, and to return the translated question followed by the four translated options. This protocol was deliberately chosen to reflect a realistic deployment in which a patient or a clinician translates a question into their own language and then poses it to the model, and it held the translation method constant across languages so that any observed difference could be attributed to the language rather than to the translation procedure.

The translations were generated once and reused for the evaluation, so that the evaluation in each language used an identical translated text, and no question was translated twice for the same language, which would have introduced an avoidable source of variation across the evaluation calls.

Evaluation protocol

The model was asked to answer each question in English and then in each of the six translated languages, and its answer was compared against the gold answer recorded in the MedMCQA dataset. The evaluation prompt asked the model to choose the single best option and to output only the letter A, B, C or D, and the first of those letters appearing in the model's response was taken as its answer, so that a response that reasoned at length and then stated an answer was scored on the stated answer rather than on the reasoning.

The model was queried at a fixed temperature of zero, so that the evaluation was deterministic and the results could be reproduced by re-running exactly the same calls, and the same maximum response length was applied to every call, so that the reasoning budget did not differ across languages and could not itself produce a disparity. Accuracy was computed per language as the proportion of the 60 questions answered correctly, and the English-to-language gap was computed as the English accuracy minus the language accuracy in percentage points, so that a positive gap indicates a language that performed worse than the English reference.

The evaluation used the DeepSeek application programming interface with the deepseek-v4-pro model released for this study, a model that exposes both a final answer and an internal reasoning trace, and only the final answer itself was scored, so that the evaluation reflects the answer a user would actually receive. No question was answered more than once per language, so that each language's accuracy was the average of one response per question across the same 60 questions, and the evaluation therefore required 420 answer calls in addition to the 360 translation calls made earlier.

Statistical analysis

The primary comparison was descriptive rather than confirmatory, because the interest lay in the size and the ranking of the accuracy gaps across languages rather than in any single hypothesis test, and the per-language accuracy was reported with its exact proportion and the corresponding gap from English. The paired design, in which the same 60 questions were used in all seven languages, meant that each language's accuracy could be compared with the English reference without the variability that would arise from different question samples, and the gap between English and each language is therefore a paired difference on identical content rather than a difference between two independent samples. Analyses were implemented in Python 3.13 with the pandas and requests libraries, and the complete script, including the translation and evaluation prompts, was archived so that the evaluation could be reproduced, with the translations stored separately so that the translation step and the evaluation step could be inspected independently and a translation-quality check could be added in future work.

RESULTS

The evaluation sample comprised 60 dental questions drawn from the 1,318 questions of the MedMCQA Dental subset, of which 40 were single-answer and 20 were multi-answer, a ratio that mirrors the composition of the full Dental subset of 1,318 questions (Table 1). The same 60 questions were posed in English and in the six translated languages, so that every accuracy reported below is computed on the identical question set, and the only thing that changes from one language to the next is the language of the question and its options.

Table 1. Characteristics of the evaluation sample.

Characteristic	Value
Total dental questions	1318
Sampled questions	60
Single-answer questions	40
Multi-answer questions	20

Note: The sample was drawn with a fixed seed from the 1,318 dental questions of the MedMCQA validation split.

Reference language

In English the model answered 61.7% of the 60 dental questions correctly, which served as the reference against which every one of the six translated languages was compared (Table 2). This reference accuracy is the value from which the language gaps reported below are measured, and it is the accuracy that a dental evaluation conducted in English would report as the model's performance, with no indication of how that performance would change in another language.

European languages

Table 2. Accuracy and English-to-language gap by language.

Language	Accuracy (%)	Gap vs English (%)
English	61.7	0.0
Swahili	26.7	35.0
Hindi	55.0	6.7
Arabic	58.3	3.3
Chinese	60.0	1.7
Portuguese	61.7	0.0
Spanish	58.3	3.3

Note: Gap is English accuracy minus language accuracy in percentage points. Portuguese accuracy equalled English.

The two European languages held closest to the English reference, and between them the Romance language group showed almost no penalty. Portuguese reached 61.7%, identical to English and therefore a gap of 0.0 percentage points, and Spanish reached 58.3%, a gap of 3.3 percentage points (Table 2).

The accuracy of the model therefore persisted almost unchanged when the questions moved into the Romance languages, and the English-to-Portuguese gap of zero indicates that translation into Portuguese introduced no measurable loss in this sample, while the small Spanish gap is within the sampling uncertainty of the 60-question set. The European block as a whole confirmed that the model's dental knowledge was intact when the question was expressed in a closely related high-resource language.

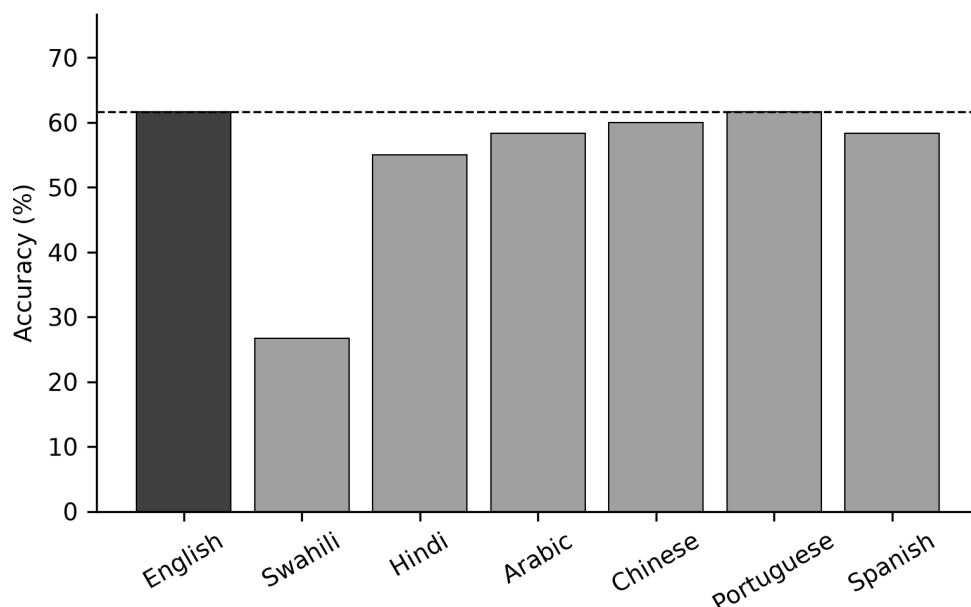
Asian languages

The three Asian languages diverged from the reference to differing degrees. Chinese reached 60.0%, a gap of 1.7 percentage points, and Arabic reached 58.3%, a gap of 3.3 percentage points, whereas Hindi fell further to 55.0%, a gap of 6.7 percentage points (Table 2). Between the closest Asian language, Chinese, and the farthest, Hindi, the accuracy gap widened by 5.0 full percentage points, so the Asian languages were not a uniform block but a graded sequence in which the penalty grew from Chinese to Arabic to Hindi, mirroring the relative abundance of medical-domain text available in each language.

African language

Swahili, the single African language in the comparison, produced the largest and most abrupt decline in the trajectory. The model answered 26.7% of the questions correctly in Swahili, a drop of 35.0 percentage points from the English reference, which reduced the accuracy to less than half of the reference value and close to the level of a random guess on a four-option question (Table 2). The gap in Swahili was more than five times larger than the next-largest gap, the 6.7 percentage points observed in Hindi, so the trajectory ended in a language whose performance fell far below every other language in the comparison. The transition from Hindi to Swahili was therefore the single inflection point of the whole trajectory, the point at which a gradual decline became a collapse and the model's dental knowledge effectively ceased to transfer from English (Figure 1).

Figure 1. Accuracy by language with the English reference marked.



Note: The dashed line marks the English reference accuracy. The African language, Swahili, shows the largest drop.

DISCUSSION

The objective of this study was to measure the accuracy of a large language model on dental questions in English and in six translated languages and to quantify the accuracy gap between English and each translated language, and the results trace a trajectory that was flat for most languages and then collapsed for one. From the English reference of 61.7%, accuracy held within 3.3 percentage points for Portuguese, Chinese, Arabic and Spanish, and fell by only 6.7 points for Hindi, before dropping by 35.0 points for Swahili, a language in which the model answered barely more than one question in four correctly.

The trajectory therefore shows that the language disparity documented in this study is not a uniform penalty applied to every non-English language but a severe penalty concentrated in the low-resource language, with the high-resource European and Asian languages barely affected at all, and the shape of the trajectory, a long plateau followed by a single sharp drop, is itself the finding to be explained. That distinctive shape is the most important result of the study, because it contradicts the assumption, implicit in much of the multilingual evaluation literature, that a model's non-English performance declines smoothly with linguistic distance from English.

What changed and what persisted across this trajectory is best read against the now standard training-data explanation of multilingual performance. The high-resource languages, Portuguese, Spanish, Chinese and Arabic, are among the languages best represented in the multilingual corpora on which large models are trained, and their accuracy persisted at or near the English reference, which is the pattern that the survey literature on multilingual models predicts.⁽¹¹⁾

Hindi, which is well represented but with less medical-domain text, showed a moderate penalty, and Swahili, a language with far less presence in the pretraining data, collapsed, so the size of the accuracy gap tracked the presumed abundance of training data in each language.⁽¹²⁾ The ordering of the gaps is therefore not arbitrary but monotonic with respect to a rough ranking of language resources, which is the signature of a model whose multilingual ability is a function of the data it has seen rather than of any intrinsic property of the languages themselves, and it predicts that other low-resource dental languages, beyond the single African language examined here, would show penalties comparable to the Swahili collapse. This is consistent with the finding that language models inherit the linguistic skew of their training corpora and reproduce it as a performance disparity across languages, and it extends that finding to dental question answering specifically.⁽¹⁷⁾ The persistence of the high-resource languages at or near the English reference also shows that the translation step itself did not introduce a uniform penalty, because if translation were the sole cause of the decline, the Romance languages, which are closest to English, would have suffered least, and the pattern would not have been as sharply concentrated in Swahili as it actually was.

The comparison with the multilingual medical literature places the result firmly in context. Bang et al.⁽¹³⁾ reported that a general chatbot's performance varied across languages on broad tasks, and the medical literature has since documented language-dependent accuracy on diagnostic questions, with studies of English and Turkish showing measurable differences between the two.^(14,22)

The same pattern appears in language-specific evaluations of models on traditional Chinese medicine, on Japanese radiology report structuring and on bilingual medical question answering, each of which reports a gap between the language of training and the language of evaluation, and each of which is limited to one or two languages rather than the seven traced here.⁽²³⁻²⁵⁾

The present result is sharper in one respect: by holding the questions and the translation protocol constant and varying only the language, it isolates the language effect from the many other confounds that cloud cross-study comparisons, and it shows that the effect can be negligible for some languages and catastrophic for another within a single model and a single question set.

The finding that Portuguese equalled English while Swahili collapsed by 35 full points is a concrete demonstration that a model's multilingual ability is not a single scalar but a spectrum whose extremes can be far apart, and it quantifies for dental questions what the multilingual surveys describe in aggregate.^(11,26) Where the prior multilingual medical studies compared two languages and found a modest difference, this study compared seven and found that the modest differences at one end of the spectrum are dwarfed by the catastrophic difference at the other, so the number of languages examined determines the shape of the conclusion.

The dental-specific literature on language models has so far concentrated almost exclusively on English and a few high-resource languages, and it has not examined the low-resource end of the spectrum at all. Studies of dental implantology, dental licensing examinations and dental education have compared models in English, Chinese and a handful of other well-resourced languages,

and they have reported competence without reporting how that competence degrades when the same questions are posed in a language with less training data.(6,7,19,20) The scoping review of language models in dental education, in particular, notes applications across settings but does not foreground the language question, and the implantology evaluations that do compare languages restrict the comparison to high-resource pairs such as English and Chinese.(19,20)

The present result therefore fills a gap in that literature by showing that a dental model that performs acceptably in English may perform at chance or below in a language that many dental patients actually speak, and that the disparity is invisible to an English-only evaluation. It also supplies the missing low-resource data point that the existing multilingual evaluations, all of which stop short of a genuinely low-resource language, cannot provide, and it demonstrates that the choice of languages in a multilingual evaluation determines whether the most important finding of that evaluation is visible at all. A dental evaluation that stopped at Hindi would conclude that the language penalty is small, and only the inclusion of Swahili reveals that the penalty is, for the most underrepresented language, near-total.

The health-equity implication follows directly from the trajectory. A language model that answers a dental question correctly in Portuguese but errs three times out of four in Swahili is not a neutral tool: it delivers accurate information to speakers of high-resource languages while delivering inaccurate information to speakers of a low-resource language, and in doing so it widens rather than narrows the very gap in access to oral-health information that language technology is often proposed to close.(17,27)

This matters for the deployment of dental question-answering systems in the multilingual populations of Africa, Latin America and Asia, where patients may well ask a question in a language that differs from the language in which the model was evaluated, and where a patient who receives a confident but wrong answer may act on it.(28,29) The result suggests that a model intended for such populations should be evaluated in the languages of those populations, and that an English benchmark is not a sufficient proxy for multilingual safety.(30)

The equity concern is compounded by the fact that the languages with the least training data are precisely the languages spoken by populations that already face the greatest barriers to dental care, so a model that performs worst for those populations is failing the patients who most need reliable information, and it is doing so precisely where a free, automated answer might otherwise have filled the gap left by a shortage of dental professionals.

The strengths of this study are its strictly paired design, in which the same 60 questions were used in every language so that the gaps are differences on identical content, and its fixed translation and evaluation protocol, which held temperature, prompt and response length constant across languages and therefore removed the usual confounds of cross-language comparison.

The principal limitation is the sample size of 60 questions, which makes the per-language accuracy estimates coarse and means that the smaller gaps, on the order of 1.7 to 6.7 percentage points, cannot be distinguished from noise with confidence; the 35-point Swahili gap, however, is large enough to be robust to sampling, and no plausible amount of sampling error would reduce it to the level of the other languages. A second limitation is that the translation was performed by the model under evaluation, so a poor translation could itself cause a low accuracy, and a translation-quality check would separate the model's translation ability from its answer ability; the observed pattern, in which accuracy tracked training-data abundance rather than any language-specific translation failure, argues against translation as the sole cause, but it cannot be ruled out entirely.(31)

A third limitation is that the study evaluated a single model, so the size of the disparity for other models is unknown, and the result should be read as a demonstration of the disparity's existence and shape rather than as a claim about any particular model's ranking.(14)A fourth limitation is that the evaluation was limited to multiple-choice questions, and the disparity might differ in size or shape for free-text answers, where a model can hedge or explain, so the present result is a lower bound on the gap for the kinds of open-ended dental questions that patients actually ask. These limitations bound the precision of the individual gaps but not the direction of the finding, and the central conclusion, that a large language model can perform near its English reference in several languages while collapsing in a low-resource language, does not depend on any single limitation being resolved in a particular way.

The implication for practice is that multilingual dental artificial intelligence should be validated language by language, and that the languages chosen for validation should include the low-resource languages spoken by the target population rather than only the high-resource languages that are convenient to benchmark. A deployment that reports accuracy only in English is, on the evidence of this study, reporting the most favourable language and concealing the worst-case language, and a responsible evaluation should instead report the full language spectrum and the

worst-case gap, because that worst-case gap is the number that determines whether a patient in a given language receives reliable information.

In concrete terms, a dental clinic or public-health program that plans to offer a question-answering interface in several languages should measure the model's accuracy in each of those languages on a fixed question set before launch, and should treat a large gap in any target language as a reason to restrict the interface to the languages that have been validated rather than to offer it in all of them. This is a low-cost step, because the paired evaluation used here requires only a translation and a fixed set of questions, and it would surface the kind of disparity that an English-only benchmark leaves invisible.(17,26)

For the research community, the very same method can be applied to any dental or medical question set to map the language frontier of a model before it is promoted for multilingual use, and the low-resource languages that the literature has so far ignored are precisely the ones that such a mapping should prioritize.(11,28)

The paired, translation-constant design is the methodological contribution of this study, because it converts a cross-lingual comparison, which is normally confounded by differences in the questions, the translation and the model, into a clean within-question measurement in which only the language varies. Future work can extend the same design to more models, more languages and free-text answers, and it can add a translation-quality measurement so that the translation and answer components of the gap are estimated separately, which would resolve the one residual ambiguity of the present result.(29,31)

For the developers of dental language models, the result points to two distinct remedies. The first is evaluation: a multilingual dental model should be evaluated across a language spectrum that includes a low-resource language, so that the shape of the disparity, a plateau followed by a drop, is visible rather than hidden by an English-only score.

The second is data: the collapse in Swahili is, on the training-data account, a symptom of underrepresentation that additional Swahili dental text would directly address, and cross-lingual fine-tuning of the kind explored for oncology and other specialties is a plausible route to closing the gap.(28,29) The two remedies are complementary rather than strictly alternative, because evaluation is what reveals the gap and data is what closes it, and a deployment that does neither will reproduce the disparity documented here, while a deployment that evaluates but does not collect data will at least be able to warn users honestly about the languages in which its answers cannot be trusted.

CONCLUSIONS

A large language model answered a fixed set of dental multiple-choice questions at 61.7% accuracy in English, and that accuracy persisted within 3.3 percentage points in Portuguese, Chinese, Arabic and Spanish while falling by 6.7 points in Hindi and by 35.0 points in Swahili, a drop that reduced the model to near-chance performance in that language.

The language disparity in dental question answering is therefore concentrated in the low-resource language rather than distributed uniformly across languages, and an English reference accuracy conceals a severe penalty for speakers of the least-represented language that no English benchmark would reveal.

The finding implies that dental language models intended for multilingual populations should be evaluated language by language, with the low-resource languages of the target population included explicitly, because an English benchmark is not a reliable proxy for safety in the languages where the model is most likely to fail, and the worst-case language, rather than the best, is the one that determines whether a deployment is safe.

The paired, translation-constant method used here offers a low-cost way to measure that disparity before deployment, and it should become part of the evaluation of any dental question-answering system intended for a linguistically diverse population, so that a model is never presented as capable in a language in which it has not actually been tested and never deployed in a language in which its accuracy is unknown.

REFERENCES

1. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi: 10.1038/s41586-023-06291-2
2. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding [Internet]. 2021 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/2009.03300>
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9:e45312. doi: 10.2196/45312
5. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems [Internet]. 2023 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/2303.13375>
6. Taymour N, Fouda SM, Abdelrahman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. *The Journal of Prosthetic Dentistry*. 2025;134(6):2427-2434. doi: 10.1016/j.prosdent.2024.12.016
7. Zhang J, Ubuzima P, Huang G, Lee EH, Wang Y, Xu H, et al. From accuracy to robustness: a comparative study of five advanced large language models on the Chinese dental licensing examination. *BMC Oral Health*. 2025;25(1):1975. doi: 10.1186/s12903-025-07540-8
8. Rath A. Leveraging ChatGPT to support terminology learning in oral anatomy: a mixed-methods study among linguistically diverse dental students. *BMC Med Educ*. 2025;25(1):1425. doi: 10.1186/s12909-025-07968-0
9. Wang C, Li M, He J, Wang Z, Darzi E, Chen Z, et al. A survey for large language models in biomedicine. *Artificial Intelligence in Medicine*. 2025;170:103268. doi: 10.1016/j.artmed.2025.103268
10. Zhang G, Xu Z, Jin Q, Chen F, Fang Y, Liu Y, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digit Med*. 2025;8(1):239. doi: 10.1038/s41746-025-01651-w
11. Xu Y, Hu L, Zhao J, Qiu Z, Xu K, Ye Y, et al. A survey on multilingual large language models: corpora, alignment, and bias. *Front Comput Sci*. 2025;19(11):1911362. doi: 10.1007/s11704-024-40579-4
12. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021:610-623. doi: 10.1145/3442188.3445922
13. Bang Y, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity [Internet]. 2023 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/2302.04023>
14. El Mahdaoui EH, Mykhalko YO. Language-dependent performance of large language models in medical diagnostics: A comparative study of ChatGPT-4o and Claude 3.5 Sonnet. *Polski Merkuriusz Lekarski*. 2025;53(6):821. doi: 10.36740/merkur202506117
15. Atılan AU, Çetin N. Benchmarking Large Language Models on the Turkish Dermatology Board Exam: A Comparative Multilingual Analysis. *tdd*. 2025;19(3):126-133. doi: 10.4274/tjd.galenos.2025.85856
16. Perez-Toro PA, Dineley J, Iniesta R, Zhang Y, Matcham F, Siddi S, et al. Exploring biases related to the use of large language models in a multilingual depression corpus. *Sci Rep*. 2025;15(1):36197. doi: 10.1038/s41598-025-19980-x
17. Ji Y, Ma W, Sivarajkumar S, Zhang H, Sadhu EM, Li Z, et al. Mitigating the risk of health inequity exacerbated by large language models. *npj Digit Med*. 2025;8(1):246. doi: 10.1038/s41746-025-01576-4
18. Wu Y, Zhang Y, Xu M, Jinzhi C, Xue Y, Zheng Y. Effectiveness of various general large language models in clinical consensus and case analysis in dental implantology: a comparative study. *BMC Med Inform Decis Mak*. 2025;25(1):147. doi: 10.1186/s12911-025-02972-2
19. Zhu G, Zhang X, Chen C. Assessing and enhancing the reliability of Chinese large language models in dental implantology. *BMC Oral Health*. 2025;25(1):1242. doi: 10.1186/s12903-025-06648-1
20. Kong M, Fok EHW, Yiu CKY. A Scoping Review of Large Language Models in Dental Education: Applications, Challenges, and Prospects. *International Dental Journal*. 2025;75(6):103854. doi: 10.1016/j.identj.2025.103854
21. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering [Internet]. 2022 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/2203.14371>
22. İnan B, Karadaş Ö, Odabaşı Z. Evaluation of ChatGPT-4o's Responses to Questions about Myasthenia Gravis in English and Turkish. *btd*. 2025;21(3):310-315. doi: 10.4274/bmj.galenos.2025.2025.5-9
23. Liu Y, Yuan Y, Yan K, Li Y, Sacca V, Hodges S, et al. Evaluating the role of large language models in traditional Chinese medicine diagnosis and treatment recommendations. *npj Digit Med*. 2025;8(1):466. doi: 10.1038/s41746-025-01845-2
24. Takita H, Walston SL, Mitsuyama Y, Watanabe K, Ishimaru S, Ueda D. Comparative performance of large language models in structuring head CT radiology reports: multi-institutional validation study in Japan. *Jpn J Radiol*. 2025;43(9):1445-1455. doi: 10.1007/s11604-025-01799-1
25. Sait ARW, Alkhurayif Y. Explainable Bilingual Medical-Question-Answering Model Using Ensemble Learning Technique. *Electronics*. 2025;14(20):4128. doi: 10.3390/electronics14204128
26. Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat Commun*. 2025;16(1):3280. doi: 10.1038/s41467-025-56989-2

27. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med.* 2023;183(6):589. doi: 10.1001/jamainternmed.2023.1838
28. Lee C, Kumar S, Vogt KA, Munshi M, Tallapudi P, Vogt A, et al. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. *Sci Rep.* 2025;15(1):27619. doi: 10.1038/s41598-025-10451-x
29. Rohanian M, Mehra T, Miglino N, Nooralahzadeh F, Krauthammer M, Wicki A. Towards scalable and cross-lingual specialist language models for oncology. *Sci Rep.* 2025;15(1):35480. doi: 10.1038/s41598-025-19282-2
30. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8
31. Kuligin L, Lammert J, Ostapenko A, Bressemer K, Boeker M, Tschochohei M. Prompt design for medical question answering with Large Language Models. *Machine Learning with Applications.* 2025;22:100758. doi: 10.1016/j.mlwa.2025.100758

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Maikro Osvaldo Chávez Moya, Leodanis Hernández Cabrera, Neisy Pérez Ramos.

Writing–review and editing: Maikro Osvaldo Chávez Moya, Leodanis Hernández Cabrera, Neisy Pérez Ramos.