Original Article

Benchmarking Large Language Models on the Turkish Dermatology Board Exam: A Comparative Multilingual Analysis

Ahmet Uğur Atılan, Niyazi Çetin
Department of Dermatology and Venereology, Pamukkale University Faculty of Medicine, Denizli, Türkiye

Abstract

Aim: Large language models (LLMs) are increasingly integrated into medical education; however, their performance on dermatology examinations in non-English contexts has not been extensively studied. This study aimed to evaluate the performance of six LLMs in terms of accuracy, error profile, and response time on the Turkish Dermatological Society (TDS) qualifying examination.

Materials and Methods: Two hundred publicly available multiple-choice questions from the TDS exam were submitted to six LLMs (ChatGPT-4, Gemini-2.0, Claude-3.7, Grok-3, DeepSeek-R1, Qwen-2.5). Each model was tested in Turkish and in English, under both batch and single-item prompt formats. The strengths and weaknesses of the models were tested under different conditions.

Results: Claude-3.7 and Grok-3 performed best (~83-84% correct) with low variance, whereas Qwen-2.5 and DeepSeek-R1 had lower accuracy (~75%) with more simple errors. Across all models, switching from Turkish to English increased median accuracy by 19.5% ($P = 0.028$). In contrast, batch vs. single-item prompting showed no overall performance difference ($P = 0.280$). DeepSeek-R1 was markedly slower (≥ 10 minutes per question vs ~134 seconds for others, $P < 0.001$). All models achieved high accuracy on common conditions but struggled with nuanced cases and negatively phrased questions.

Conclusion: Current LLMs can answer standard dermatology certification questions with moderate to high accuracy, especially in English. However, they are still susceptible to linguistic traps, negation, and nuanced clinical distinctions. Before they can be routinely used for educational or clinical purposes, optimization for Turkish language input and complex reasoning is necessary.

Keywords: Large language models (LLMs), artificial intelligence in dermatology, Turkish dermatology board examination, multilingual ai performance, prompt engineering

Introduction

Artificial intelligence (AI) technologies are used as an effective tool in medical education to support the acquisition of theoretical knowledge and improve the clinical skills of medical students and residents.[1] The increasing use of AI applications in medical education has potential, especially in disciplines based on visual diagnoses, such as dermatology; this trend necessitates re-evaluating conventional assessment tools, such as specialty competency exams, with AI models.[2] Studies on the performance of AI models in medical education exams reveal the potential, limitations, and room for improvement of this technology.[3] Recent benchmark studies show that state-of-the-art large language models (LLMs) (e.g., GPT-4, Gemini Advanced, Claude) can exceed the 60% pass mark on united states medical licensing examination step 1-style items and perform at or near the resident level in ophthalmology and orthopedic vignette sets.[3-5]

Despite the progress made, two critical knowledge gaps remain. First, almost all validation studies have been conducted in English, while more than half of the world's medical

Submission: 26-May-2025
Acceptance: 10-Jul-2025 **Web Publication:** 10-Sep-2025

Access this article online

Quick Response Code:

Website: www.turkjdermatol.com

DOI: 10.4274/tjd.galenos.2025.85856

Address for correspondence: Ahmet Uğur Atilan, Assoc., Prof., MD, Department of Dermatology and Venereology, Pamukkale University Faculty of Medicine, Denizli, Türkiye
Email: auatilan@pau.edu.tr
ORCID ID: 0000-0001-9244-1011

How to cite this article: Atılan AU, Çetin N. Benchmarking large language models on the Turkish dermatology board exam: a comparative multilingual analysis. Turk J Dermatol. 2025;19(3):126-133.

students study in other languages. LLMs show significant performance declines in low-resource languages due to issues such as imbalanced training data, cultural differences, and tokenization problems.[6-8] Second, there has been limited research into domain-specific, high-stakes examinations that assess nuanced clinical reasoning rather than just general medical knowledge.

The Turkish Dermatology Society (TDS) qualifying examination combines essential knowledge with image-rich clinical scenarios and is primarily administered in Turkish. To our knowledge, no published study has yet assessed contemporary LLMs using this examination, nor has any compared their performance when identical items are presented in Turkish versus professionally translated English. Addressing this gap is crucial for two main reasons. First, educators need evidence before incorporating AI into residency training. Second, algorithm developers require detailed error profiles to optimize multilingual models and prevent hallucinations or unsafe recommendations.

In this study, we conducted a comparative performance analysis of six publicly accessible LLMs using multiple-choice questions from the TDS qualifying examination. To explore the effects of linguistic and structural variations on model performance, we tested each model under four prompting conditions that varied by input language (Turkish vs. English) and delivery format (batch vs. single-item). By systematically comparing accuracy, response latency, and error characteristics, we aimed to evaluate the dermatological knowledge base as well as the language adaptability of these models. We, therefore, benchmarked six contemporary LLMs on 200 standardized text-only TDS board items to quantify language-related and prompt-related performance shifts and to characterize error profiles relevant to clinical reasoning.

Materials and Methods

Study Design

In this study, a prospective benchmarking comparing the performance of six publicly available LLMs on the dermatology specialty examination was conducted (Table 1). All analyses were performed between 15 February and 1 March 2025 to minimize version drift. Each model was initialized in a fresh "clean" account session to avoid any carryover of prior context. No plug-ins or speech memory features were activated.

Question Bank

Two hundred multiple-choice questions were selected from the publicly available repository of the TDS qualifying examination. Items based on clinical photographs or histopathology images were excluded to keep the assessment entirely text-based, and questions assessing epidemiology, pathophysiology, clinical diagnosis, and treatment were included.

Translation

The initial English drafts of all 200 Turkish board examination items were created using DeepL Pro (v3.5). Subsequently, a senior dermatology resident proficient in academic English (N.Ç.) and a professional dermatologist with a high level of proficiency in academic and clinical English (A.U.A.) reviewed the machine translations. Together, they reached a consensus, correcting any inaccuracies in medical terminology and addressing cultural nuances.

Prompting Conditions

Batch conditions involved multiple questions uploaded simultaneously, whereas single-item conditions involved uploading questions individually. During batch uploading, four separate Word files were uploaded one by one (batch Turkish 2015, batch Turkish 2017, batch English 2015, batch English 2017). In single-item (sequential) prompting, 400 questions were uploaded individually each time.

Outcome Measures

For each method, the response times and accuracy rates of the models were analyzed. The language factor was examined by averaging the batch and single-item prompting results. The official answer key was used as a reference. For each correct answer, 1 point was given and 0 points for an incorrect answer; the total number of correct answers and the success percentage of each model were calculated. Correct answer rates were reported separately for each model and method, and comparisons were made between models and methods. In addition, questions that all models answered incorrectly, questions that only one model answered correctly (superior performance), and questions that only one model answered incorrectly (simple error) were analyzed. Across all four methods, any question answered incorrectly by at least five of the six models was defined as a "difficult question". Also, all 200 questions were categorized into six content domains: (1) common dermatoses and first-line management, (2) clinical case vignettes, (3) rare syndromes and eponyms, (4) disease sub-typing, (5) negatively worded stems, and (6) other. Accuracy was subsequently assessed for each category to enable a category-based performance analysis. In the batch Turkish method, the response times of the models were determined using a stopwatch.

TABLE 1. Details of the language models used in this study, including provider platforms and access dates

Model	Provider (API/UI)	Date Accessed
ChatGPT-4.0	OpenAI (web)	18 Feb 2025
Gemini 2.0 Flash	Google DeepMind (web)	21 Feb 2025
Claude 3.7 Sonnet	Anthropic (web)	17 Feb 2025
Grok-3	xAI (web)	20 Feb 2025
DeepSeek R1	DeepSeek AI (web)	19 Feb 2025
Qwen 2.5	Alibaba (web)	20 Feb 2025

AI: Artificial intelligence, API: Application programming interface, UI: User interface

Statistical Analysis

All statistical analysis were performed using IBM SPSS Statistics v26.0 (IBM Corp., Armonk, NY) software. The distribution of continuous variables was examined using the Shapiro-Wilk test, and non-parametric tests were preferred when the normality assumption was not met. The significance level was set at $P < 0.05$ for all tests.

Language effect: The average performance of the models in Turkish (batch Turkish + single-loading Turkish) and English (batch English + single-loading English) formats was compared using the paired Wilcoxon signed-rank test.

Method effect: The effect of the batch and single-loading methods in each language group was analyzed using the paired Wilcoxon test.

Response time analysis: The average response time of the DeepSeek model was compared with the response times of other models using the Mann-Whitney U test due to the non-normal distribution of the data. The Kruskal-Wallis test was used to compare models other than DeepSeek.

Difficult questions and word count: The average number of words in difficult questions that were answered incorrectly by all models or correctly by only one model, was analyzed using the Mann-Whitney U test.

Ethical Considerations

The study analyzed publicly available examination material and generated AI responses; it involved no human participants or patient data and therefore did not require Institutional Review Board approval. All procedures conformed to the Declaration of Helsinki principles for non-interventional research.

Data Availability

Full prompt templates, anonymized model outputs, and analysis scripts can be requested from the corresponding author if needed.

Results

Overall Performance Evaluation by Model

When analyzing the overall number of correct answers and average performance of the models using the "batch Turkish" and "single-loading Turkish" methods, the Claude (84.0%±0.00) and Grok-3 (83.0%±3.83) models demonstrated the most successful results, showing the highest average number of correct answers and low standard deviations. In contrast, the Qwen 2.5 (74.25±613) and DeepSeek (75.5%±7.05) models displayed the lowest performance and highest inconsistency, indicated by both lower average correct answer counts, and particularly, for DeepSeek, higher standard deviation values (Figure 1).

However, some models demonstrated superior performance in specific domains. For example, among the 200 questions, there were items that only Qwen 2.5 answered correctly while all other models failed, suggesting areas where it outperformed its peers.

Impact of Language Factor

There was a significant performance advantage for English versus Turkish prompts across models ($P = 0.028$). This result indicates that LLMs perform significantly better in English than in Turkish. A noticeable performance improvement was observed across all models when switching to English questions (Figure 2). In particular, the ChatGPT and Qwen 2.5 models were the most positively impacted by the language change.

Effect of Method Factor

The DeepSeek model demonstrated a significant performance improvement in the Turkish single-item prompting method compared to the Turkish batch method, showing the largest gain from this approach. In contrast, the single-loading method led to a performance decline in the ChatGPT and

Gemini models. When comparing batch and single-loading methods, no statistically significant difference was observed between the methods ($P > 0.05$). However, per-model analyses revealed notable individual differences beyond this general finding (Figure 3).

Simple Errors and Inconsistencies

In the analysis of "simple errors," the DeepSeek and Qwen 2.5 models had the highest number of simple errors. DeepSeek recorded the most errors with a total of 27, followed by Qwen with 22 (Figure 4a). When comparing prompting conditions, the Turkish batch condition yielded the most errors, whereas the English single-item condition had the fewest (Figure 4b).

Word Count Analysis of Difficult Questions

The difficult questions had a significantly lower average word count than other questions (9.71±7.08 vs. 11.9±9.89 words; $P = 0.019$), suggesting that shorter questions tended to pose more of a challenge (Figure 5).

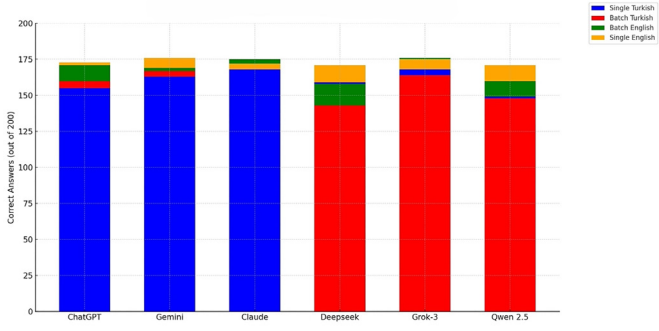

Figure 1. Prompting condition. Stacked segments represent the number of questions each model answered correctly in four prompting conditions: single-item Turkish (blue), batch Turkish (red), batch English (green) and single-item English (orange)

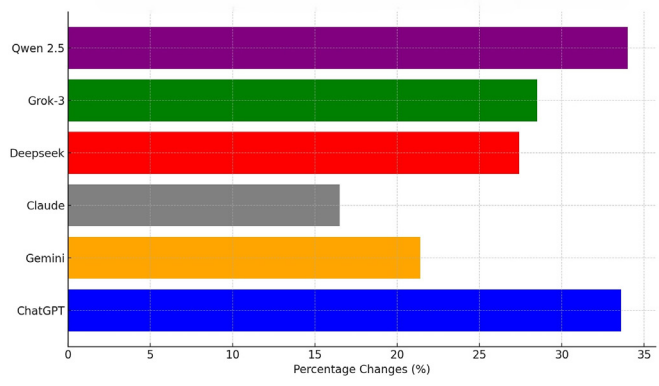

Figure 2. English versus Turkish condition. Horizontal bars show the absolute percentage-point reduction in model error rates when item prompts are translated into English. A positive value indicates higher accuracy in English

Category-Based Performance Evaluation

When evaluating the full set of results across 200 questions, six language models and four prompting methods, the models achieved over 90% accuracy in most categories, including common disease presentations, primary diagnoses, and

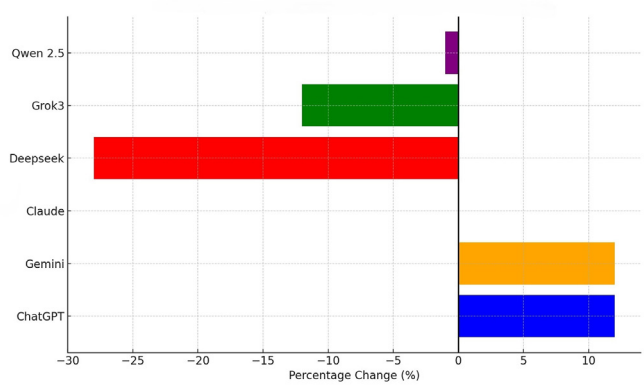

Figure 3. Batch versus single-loading condition. Bars show the percentage-point difference in error rates when models were prompted in batch versus single-item format. Negative values indicate that the model made fewer errors when given one item at a time

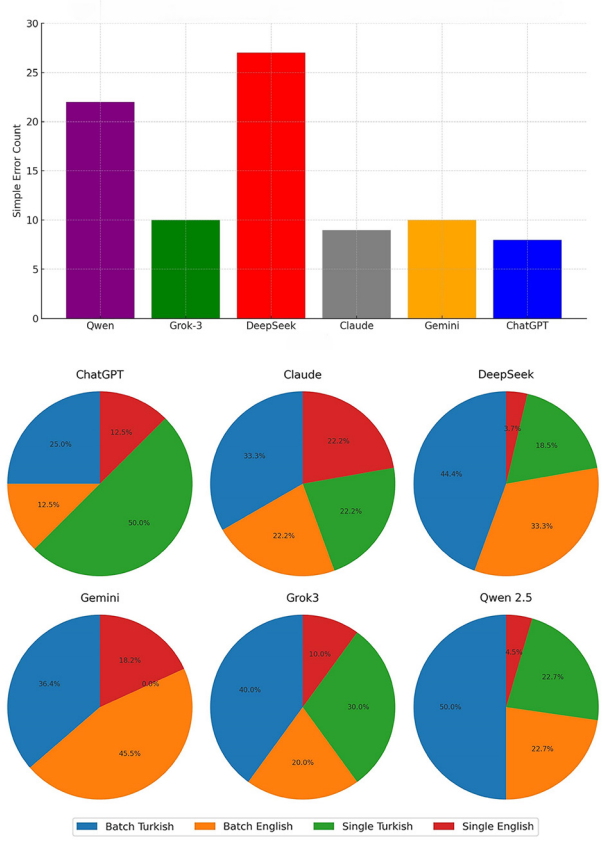

Figure 4. Simple error distribution by model and prompting condition. (a) Total number of simple errors made by each model across all tasks (b) Proportional distribution of simple errors per model under four prompting conditions: single-item Turkish, batch Turkish, batch English, and single-item English

treatment approaches, clinical case scenarios, and questions involving specific medical terminology or rare syndromes. In contrast, the lowest accuracy rates were observed in distinguishing clinical subtypes (57.14%) and in handling negatively phrased questions (83.33%) (Table 2).

Response Time: Performance or Speed?

DeepSeek demonstrated a significantly longer response time, consistently exceeding 10 minutes per item (720 seconds), while the other five models responded within a comparable time frame (mean: 134±73.85 seconds), with no statistically significant differences observed among them ($P > 0.05$) (Table 3).

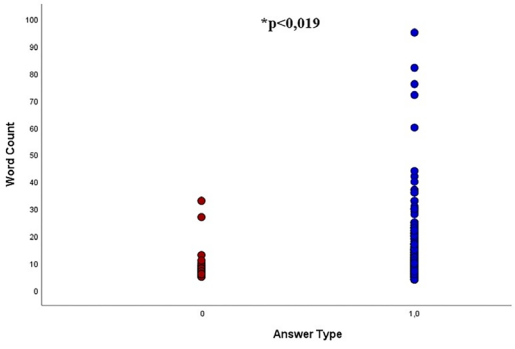

Figure 5. Word count and accuracy relation. Each point represents a board question, plotted by word count and whether it was answered correctly (1) or not (0). Incorrectly answered questions had significantly shorter text length ($P < 0.019$)

Discussion

This study presents the first direct comparison of six contemporary LLMs using the TDS qualifying examination. Three key findings stand out. First, the language used was the primary factor influencing performance: switching from Turkish to English improved the median accuracy significantly, benefiting all tested LLMs. Second, the presentation method had minimal overall impact; however, DeepSeek R1 performed significantly better with single-item prompts. Third, even the best-performing models faced challenges with nuanced clinical differentiation, negatively worded questions, and time efficiency. This highlights ongoing limitations in contextual reasoning and practical usability.

The results of the overall performance evaluation showed that LLMs have gained significant competence in interpreting and applying medical knowledge. The consistent performance of the Claude and Grok-3 models, suggests that these models have a more balanced information processing capacity. Claude has shown successful performance in studies evaluating LLMs.[9] In radiology board exams, Claude outperformed Bard and Gemini Pro by achieving 62% accuracy.[9] In NBME exams, it again performed similarly to GPT-3.5 and Bard with a score of 84.7%.[10] Grok 3, on the other hand, is still under development, and while it shows potential in interaction skills and mathematical reasoning, its performance in medical exams has not yet been extensively evaluated.[11] As both models continue to evolve, their role in medical education and examinations will likely expand, and they will need to be

TABLE 2. Accuracy of language models across predefined dermatology question categories (category-based analysis)

Category	Items (n)	Accuracy (%) ± SD
Common dermatoses / first-line management	35	97.1 ± 16.5
Clinical case vignettes	27	94.4 ± 8.0
Rare syndromes / eponyms	22	95.5 ± 10.0
Disease sub-typing	21	57.1 ± 1.0
Negatively worded stems	12	83.3 ± 20.0

SD: Standard deviation

TABLE 3. Response times and performance characteristics of each language model

Model	Response Time	Description
ChatGPT 4.0 (OpenAI)	80 seconds	Fast and stable
Gemini 2.0 Flash (Google DeepMind)	50 seconds	The fastest responding model
DeepSeek R1 (DeepSeek AI)	>10 minutes	Extremely slow / server congestion
Grok 3 (xAI)	180 seconds	Moderate waiting time
Claude 3.7 Sonnet (Anthropic)	230 seconds	Slow responding model
Qwen 2.5 (Alibaba)	120 seconds	Balanced response time / Medium speed

AI: Artificial intelligence

regularly re-evaluated and refined to ensure their reliability and relevance in the field.[11] On the other hand, Qwen 2.5 and DeepSeek's fluctuating performance and susceptibility to simple errors reflect differences in model architectures and training strategies.[12,13]

It was observed that the language factor had a significant effect on the AI models. The significantly higher performance of the models on English questions compared to Turkish questions reveals the dominance of English data sets in the training processes of LLMs.[14,15] This aligns with reports in the literature that LLMs perform worse in languages other than English.[6,7] LLMs are more successful in English in part because of the vast amount of English digital content and the concentration of AI research on English, owing to that language's global dominance.[15-17] Non-English languages present unique challenges (e.g., cultural nuances, complex linguistics) that require specialized AI approaches. A lack of standardized resources and tools in these languages, can lead to issues like cultural hallucinations, making it more difficult to develop effective AI models for them.[8] Despite English's privileged position in AI development, there is growing recognition of the need to improve LLM performance in other languages. Initiatives like cross-language training and multilingual model development are working to create more inclusive, culturally sensitive AI systems.[14,18]

Although there was no statistically significant difference between batch and single-item prompting in the analyses regarding the method factor, model-based differences are noteworthy. The performance improvement of the DeepSeek model in the Turkish single-item prompting method suggests that some models are more sensitive to sequential processing.[19] AI systems designed for sequential processing use character recognition, on-the-fly verification, and error correction mechanisms to ensure accuracy during real-time data entry.[19,20] These approaches provide high accuracy and user efficiency by reducing errors in data entry.[19] This finding suggests that the prompt dependency and context management capabilities of LLMs may vary from model to model.[21] Unlike DeepSeek, models like ChatGPT and Claude experienced a decline in performance under the same conditions, underscoring the importance of tailoring LLM deployment strategies to model-specific strengths and intended use cases.

In particular, category-based analyses clearly revealed the strengths and weaknesses of AI models. High success rates in basic medical knowledge and common conditions confirm the potential of these models to provide knowledge-based support in general medical practice.[22] However, high error rates in distinguishing clinical subtypes of diseases and negatively worded question stems suggest that AI models still have limitations in analyzing context in depth and overcoming linguistic pitfalls.[23-25]

This finding is in line with the known difficulties of negation and contextual disambiguation in natural language processing systems.[26,27] Moreover, the questions that stumped all models were notably short, suggesting that LLMs make more errors on context-free, brief, and ambiguous statements. As previous studies also suggest, LLMs are heavily context-driven, and their performance degrades when information is lacking.[28]

In terms of response times, the trade-off between speed and another aspect of performance must also be considered. For AI systems used especially in clinical applications, not only accuracy but also speed in practical use is critical.[29]

From an educational perspective, LLMs already demonstrate near-expert-level performance on routine factual dermatology content and could be useful as supplementary tutoring tools, particularly when prompts are provided in English. However, their susceptibility to short, context-poor questions and semantic traps presents a risk if they are used uncritically for high-stakes self-assessment. Moreover, DeepSeek R1's extremely long response time (over 10 minutes per question) makes real-time feedback impractical.

Study Limitations

Several limitations should be taken into account when interpreting our findings. First, our analysis was limited to 200 publicly available, text-only multiple-choice items. This excluded image-based and open-ended questions, which are essential in dermatology practice, therefore, the model's performance on multimodal or free-text tasks remains unassessed. Second, due to the rapid development of LLMs architectures and public interfaces, our results reflect the model versions as of February 2025 and may not apply to future iterations. Third, all assessment items were derived from a single national board examination, which restricts the external validity to other dermatology curricula or broader medical fields. Lastly, we used a binary scoring approach, giving credit only for fully correct responses. This approach may underestimate partial reasoning or nuanced understanding that could be better evaluated using a rubric-based scoring system. Addressing these limitations will require larger, multimodal test sets, ongoing reassessment of evolving model versions, and the integration of more detailed qualitative scoring frameworks.

Conclusion

In conclusion, this study demonstrated the potential and current limitations of AI models in medical education and assessment processes from a multidimensional perspective Our findings indicate that, while AI systems can be valuable tools for medical decision support, they still require improvement in

areas such as linguistic diversity, contextual analysis, and use-case optimization. Future research should focus on developing multilingual and culturally sensitive models, enhancing context management capabilities, and optimizing the speed-accuracy balance, particularly in clinical applications.

Ethics

Ethics Committee Approval: Not applicable.

Informed Consent: Not applicable.

Footnotes

Authorship Contributions

Concept: A.U.A., Design: A.U.A., N.Ç, Data Collection or Processing: A.U.A., N.Ç, Analysis or Interpretation: A.U.A., N.Ç, Literature Search: A.U.A., N.Ç, Writing: A.U.A., N.Ç.

Conflict of Interest: The authors declared that they have no conflict of interest.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Bhuyan SS, Sateesh V, Mukul N, Galvankar A, Mahmood A, Nauman M, Rai A, Bordoloi K, Basu U, Samuel J. Generative artificial intelligence use in healthcare: opportunities for clinical excellence and administrative efficiency. J Med Syst. 2025;49(1):10.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118.
3. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on usmle step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. Med Sci Educ. 2024;34(1):145-152.
4. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, Mimouni M, Plopsky G. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. 2025;263:527-536.
5. Fabijan A, Zawadzka-Fabijan A, Fabijan R, Zakrzewski K, Nowosławska E, Polis B. Assessing the accuracy of artificial intelligence models in scoliosis classification and suggested therapeutic approaches. J Clin Med. 2024;13(14):4013.
6. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, Al-Balwah MY, Barakat M, Egger J. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infect Dis. 2024;24(1):799.
7. Ahuja K, Diddee H, Hada R, Ochieng M, Ramesh K, Jain P, Nambi A, Ganu T, Segal S, Axmed M, Bali K, Sitaram S. Mega: multilingual evaluation of generative ai. arXiv preprint arXiv:230312528. 2023.
8. Sato K, Kaneko H, Fujimura M. Reducing cultural hallucination in non-english languages via prompt engineering for large language models. OSF Preprints. 2024;10:1-8.
9. Wei B. Performance evaluation and implications of large language models in radiology board exams: prospective comparative analysis. JMIR Med Educ. 2025;11:e64284.
10. Abbas A, Rehman MS, Rehman SS. comparing the performance of popular large language models on the national board of medical examiners sample questions. Cureus. 2024;16(3):e55991.
11. Wangsa K, Karim S, Gide E, Elkhodr M. A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: ChatGPT, Bard, Llama, Ernie, and Grok. Future Internet. 2024;16(7):219.
12. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F, Hui B, Ji L, Li M, Lin J, Lin R, Liu D, Liu G, Lu C, Lu K, Ma J, Men R, Ren X, Ren X, Tan C, Tan S, Tu J, Wang P, Wang S, Wang W, Wu S, Xu B, Xu J, Yang A, Yang H, Yang J, Yang S, Yao Y, Yu B, Yuan H, Yuan Z, Zhang J, Zhang X, Zhang Y, Zhang Z, Zhou C, Zhou J, Zhou X, Zhu T. Qwen technical report. arXiv preprint arXiv:230916609. 2023.
13. Liu A, Feng B, Wang B, Wang B, Liu B, Zhao C, Zhao C, Dengr C, Ruan C, Dai D, Guo D, Yang D, Chen D, Ji D, Li E, Lin F, Luo F, Hao G, Chen G, Li G, Zhang H, Xu H, Yang H, Zhang H, Ding H, Xin H, Gao H, Li H, Qu H, Cai JL, Liang J, Guo J, Ni J, Li J, Chen J, Yuan J, Qiu J, Song J, Dong K, Gao K, Guan K, Wang L, Zhang L, Xu L, Xia L, Zhao L, Zhang L, Li M, Wang M, Zhang M, Zhang M, Tang M, Li M, Tian N, Huang P, Wang P, Zhang P, Zhu Q, Chen Q, Du Q, Chen RJ, Jin RL, Ge R, Pan R, Xu R, Chen R, Li SS, Lu S, Zhou S, Chen S, Wu S, Ye S, Ma S, Wang S, Zhou S, Yu S, Zhou S, Zheng S, Wang T, Pei T, Yuan T, Sun T, Xiao WL, Zeng W, An W, Liu W, Liang W, Gao W, Zhang W, Li XQ, Jin X, Wang X, Bi X, Liu X, Wang X, Shen X, Chen X, Chen X, Nie X, Sun X, Wang X, Liu X, Xie X, Yu X, Song X, Zhou X, Yang X, Lu X, Su X, Wu Y, Li YK, Wei YX, Zhu YX, Xu Y, Huang Y, Li Y, Zhao Y, Sun Y, Li Y, Wang Y, Zheng Y, Zhang Y, Xiong Y, Zhao Y, He Y, Tang Y, Piao Y, Dong Y, Tan Y, Liu Y, Wang Y, Guo Y, Zhu Y, Wang Y, Zou Y, Zha Y, Ma Y, Yan Y, You Y, Liu Y, Ren ZZ, Ren Z, Sha Z, Fu Z, Huang Z, Zhang Z, Xie Z, Hao Z, Shao Z, Wen Z, Xu Z, Zhang Z, Li Z, Wang Z, Gu Z, Li Z, Xie Z. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:240504434. 2024.
14. Zhu W, Lv Y, Dong Q, Yuan F, Xu J, Huang S, Kong L, Chen J, Li L. Extrapolating large language models to non-english by aligning languages. arXiv preprint arXiv:230804948. 2023.
15. Tran K. From english to foreign languages: Transferring pre-trained language models. arXiv preprint arXiv:200207306. 2020.
16. Crompton H, Edmett A, Ichaporia N, Burke D. AI and English language teaching: affordances and challenges. Br J Educ Technol. 2024;55:2503-2529.
17. Alsaedi B. Ai in learning english: enhancing english skills through the use of ai-powered language learning tools. International Journal of Education and Social Science Research. 2024;7(5):82-94.
18. Mujadia V, Urlana A, Bhaskar Y, Pavani PA, Shravya K, Krishnamurthy P, Sharma DM. Assessing translation capabilities of large language models involving English and Indian languages. arXiv preprint arXiv:231109216. 2023.
19. Furuhata Y. Climatic media: Transpacific experiments in atmospheric control: Duke University Press; 2022.
20. Breure L. Interactive data entry: problems, models, solutions. History and Computing. 1995;7:30-49.
21. Palaniappan R, Surendran S, Balakrishnan S. Enhanced U-Net with transformer-driven encoder for medical image segmentation. 2024 International Conference on Brain Computer Interface and Healthcare Technologies (iCon-BCIHT), 1-8.
22. Liu j, Cai W, Yang L, Zhao S, Wu C, Chen L, Chang X, Yang Y, Xing L, Liang X. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. arXiv preprint arXiv:230414204. 2023.
23. Koga S. Exploring the pitfalls of large language models: inconsistency and inaccuracy in answering pathology board examination-style questions. Pathol Int. 2023;73(12):618-620.
24. Oeding JF. Editorial commentary: studies evaluating artificial intelligence large language models' ability to respond to questions are repetitive and out-of-date: artificial intelligence must now be applied to improve clinical practice and patient care. Arthroscopy. 2025;41(6):2009-2011.

25. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Cai ZR, Daneshjou R, Rajpurkar P. Testing the limits of language models: a conversational framework for medical AI assessment. medRxiv. 2023.
26. Zou B, Zhu Q, Zhou G. Negation focus identification with contextual discourse information. proceedings of the 52nd annual meeting of the association for computational linguistics. Baltimore, Maryland, USA; 2014:522-530.
27. Yadav P, Kashyap I, Bhati BS. Contextual ambiguity framework for enhanced sentiment analysis. Tehnički Glasnik. 2024;19(3):385-393.
28. Cao B, Cai D, Zhang Z, Zou Y, Lam W. On the worst prompt performance of large language models. arXiv preprint arXiv:240610248. 2024.
29. Madhavi JS, Agrawal OD. Potential use of artificial intelligence in a healthcare system. The Chinese Journal of Artificial Intelligence. 2022;1:e050822207306.