TMSJ
TURKISH MEDICAL STUDENT JOURNAL

Turk Med Stud J 2025;12(3):60-7
DOI: 10.4274/tmsj.galenos.2025.2025-7-4

ORIGINAL ARTICLE

ASSESSING THE PERFORMANCE OF WIDELY USED LARGE LANGUAGE MODELS ACROSS MEDICAL DISCIPLINES USING USMLE-STYLE EXAM QUESTIONS: AN IN-DEPTH EVALUATION

Zeynep Serra Özler[1], Betin Bilkan Karaman[1], Eray Atalay[2]

[1]Eskişehir Osmangazi University School of Medicine, Eskişehir, TÜRKİYE
[2]Eskişehir Osmangazi University School of Medicine, Department of Ophthalmology, Eskişehir, TÜRKİYE

ABSTRACT

Aims: Large language models are increasingly used in medical education and clinical decision-making. While previous studies have demonstrated that individual large language models can perform well on standardized medical exams, comparative evaluations across multiple large language models and medical disciplines remain limited. This study aimed to evaluate and compare the performance of seven large language models-generative pre-trained transformer-4o, DeepSeek-R1, DeepSeek-V3, Llama 3.3, Gemini 2.0 Flash, Claude 3.7 Sonnet, and OpenBioLLM on United States Medical Licensing Examination -style multiple- choice questions.

Methods: A total of 1000 questions were randomly selected from 25 medical disciplines from AMBOSS question-bank, excluding those with images, tables or charts. Each model was prompted with a standardized system and user instruction designed to produce a single letter answer without explanation. Evaluations were conducted across three independent runs per model using a temperature of 0.0; for models supporting seed control, predetermined seeds were used to ensure reproducibility. Version identifiers and access dates were documented to ensure reproducibility.

Results: Generative pre-trained transformer-4o achieved the highest accuracy (89.3%), followed by DeepSeek-R1 (87.0%) and Llama 3.3 (84.1%), while OpenBioLLM and DeepSeek-V3 scored the lowest (78.2% and 76.5%, respectively). Generative Pre-Trained Transformer-4o led in 14 of 25 disciplines, especially clinical ones, while DeepSeek-R1 excelled in public health-oriented subjects. Performance varied significantly across disciplines, with infectious diseases (91.4%), psychiatry (91.1%), and behavioral science (89.3%) showing the highest scores, while cardiology (67.5%) and genetics (76.1%) were the most challenging areas.

Conclusion: Generative pre-trained transformer-4o and DeepSeek-R1 outperformed other models across a wide range of medical disciplines. However, substantial variability across disciplines and models highlights current limitations in large language model reasoning, particularly in complex fields like cardiology. While these findings highlight the potential of large language models in medical education, further development and rigorous validation are required before they can be reliably integrated into clinical practice and medical education.

Keywords: Artificial intelligence, large language models, medical education

INTRODUCTION

Large language models (LLMs) are becoming essential tools across numerous fields, including medicine (1). Initially, LLMs were primarily developed by major technology companies using proprietary, closed-source frameworks, such as OpenAI's generative pre-trained transformer (GPT) series and Google AI's Gemini. However, the emergence of open-source LLMs is reshaping the field by expanding accessibility and flexibility, and creating new opportunities, particularly in the medical field.

The potential applications of such tools in medical education and clinical practice are being increasingly explored and their scope is expanding to address the needs of a broad audience ranging from medical students to experienced healthcare providers (2). As such, evaluating the performance of LLMs in medical knowledge assessment has become a key area of research interest, with

Address for Correspondence: Zeynep Serra Özler, Eskişehir Osmangazi University School of Medicine, Eskişehir, TÜRKİYE
e-mail: ozlerzeynepserra@gmail.com
ORCID iD of the authors: ZSÖ: 0000-0003-0013-1230; BBK: 0000-0001-6240-9651; EA: 0000-0002-2536-4279
Received: 31.07.2025 Accepted: 29.09.2025 Publication Date: 27.10.2025

Cite this article as: Özler ZS, Karaman BB, Atalay E. Assessing the performance of widely used large language models across medical disciplines using usmle-style exam questions: an in-depth evaluation. Turk Med Stud J. 2025;12(3):60-7.

numerous studies analyzing their ability to accurately answer questions from standardized medical exams to third-party question banks (3, 4).

Given the complex nature of questions used in medical exams, which requires both the ability to apply medical knowledge and clinical reasoning in real-world scenarios, medical students often refer to third-party resources including LLMs such as ChatGPT, DeepSeek and others (5). Notably, ChatGPT has been shown to achieve scores above the required threshold for Step 1, Step 2 clinical knowledge, and Step 3 United States Medical Licensing Examination (USMLE) exams (6). Recent research has also shown that DeepSeek-R1 demonstrates medical reasoning capabilities, suggesting its promising role in medical education and clinical decision-making (7). However, the accuracy of these tools may vary across disciplines, performing well in certain disciplines while generating false interpretations and reasonings in others.

Although previous research has demonstrated that individual LLMs can successfully pass specific medical licensing exams (8, 9), there is a lack of studies that compare the performance of the latest LLMs across different disciplines of medicine. In this study, we aim to assess the performance of multiple LLMs, including both proprietary and open-source models, in answering USMLE-style questions derived from AMBOSS, a third-party USMLE-style question-bank, covering both preclinical and clinical medical disciplines.

MATERIALS AND METHODS

This study did not require research ethics approval as it did not involve human subjects. To compare the performance of various LLMs, the study utilized 1000 USMLE-style multiple-choice questions (MCQs) sourced from AMBOSS (10), a non-public widely used medical education platform with a comprehensive question bank, to prevent learning effects and eliminate bias from publicly accessible question sets. To ensure diversity across different disciplines, 40 text based questions were randomly selected using a random number generator from each of the 25 medical disciplines (allergy and immunology, anatomy and embryology, behavioral science, biochemistry, biostatistics and epidemiology, cardiology, endocrinology, gastroenterology, genetics, hematology, histology and molecular biology, infectious diseases, legal medicine and ethics, microbiology, nephrology, neurology, obstetrics and gynecology, pathology, pediatrics, physiology, psychiatry, public health, pulmonology, rheumatology, and surgery) across different blocks. To ensure compatibility with LLM interfaces, questions that included images, charts, or tables were excluded. The final dataset included the question stem, five answer options (A-E), the correct answer (ground truth), and the corresponding category label. The question set likely reflects Step 1 content, though difficulty level was not formally stratified.

Seven LLMs were evaluated in this study (Supplementary Material S1). GPT-4o was accessed via the official OpenAI application programming interface (API) on March 13, 2025. Claude 3.7 Sonnet was accessed on March 13, 2025, and Gemini 2.0 Flash on March 15, 2025, both via their respective official APIs. Llama 3.3 70B was accessed through the Groq API on March 14, 2025. OpenBioLLM 70B, DeepSeek-V3, and DeepSeek-R1 were accessed via the Nebius API on March 19, 2025. These version identifiers and access dates were documented to ensure full transparency and reproducibility, as LLM capabilities may evolve over time with ongoing model updates. The models were used with their default parameters as provided by the official APIs, without further optimization or fine-tuning.

Each model received a standardized prompt comprising a system-level instruction and a user-level message. The system prompt instructed the model to act as a highly knowledgeable medical expert with extensive experience in clinical reasoning and to select the most evidence-based and clinically appropriate answer without explanation. The user prompt presented the question stem followed by the five answer choices labeled A-E and instructed the model to respond with only a single uppercase letter corresponding to its answer, without any punctuation or explanation. This prompt was applied uniformly across all runs and models.

Each model was evaluated across three independent runs to assess the consistency of performance. For models that support deterministic outputs via seed control (GPT-4o, Gemini 1.5 Flash, Llama 3.3 70B, DeepSeek V3, DeepSeek R1, and OpenBioLLM 70B), distinct predetermined random seeds were used for each run as recommended in recent work on reproducible LLM evaluation (11). A random seed serves as a fixed numerical starting point that regulates the model's internal randomization; by fixing the seed, the same input under the same conditions is expected to produce the same output, thereby enabling reproducibility. Varying the seed across runs allowed evaluation of performance under controlled, replicable conditions. The Claude 3.7 Sonnet model does not currently support seed control; hence, its responses were treated as stochastic across trials.

The temperature parameter was set to 0.0 for all models. In LLMs, temperature is a hyperparameter that influences the probability distribution used during text generation: higher temperatures increase variability by allowing the model to select less likely tokens, while lower temperatures narrow the distribution, producing more focused and deterministic outputs. Setting the temperature to 0.0 effectively eliminates randomness in token selection. This forces the model to consistently choose the most probable next token at each step, ensuring stable outputs across runs (12).

Output post-processing was minimal; however, for DeepSeek models, structured reasoning tags (e.g., <THINK>) were removed to isolate the final answer selection. No additional preprocessing was applied to the output of other models.

Statistical Analysis

All analyses were conducted in R (version 4.2.2; R Foundation for Statistical Computing, Vienna, Austria). Accuracy was defined

as the proportion of correct responses for each of the seven language models. To assess whether overall accuracy differed among models, a global chi-square test of independence was performed on the 7×2 contingency table of model by response correctness. Upon obtaining a significant global χ^2 result (α=0.05), pairwise comparisons of proportions between every pair of models were carried out using two-sided chi-square tests. P-values below 0.05 were considered statistically significant.

RESULTS

A total of 1000 MCQs from 25 medical disciplines were administered to seven LLMs: GPT-4o, DeepSeek-R1, DeepSeek-V3, Llama 3.3, Gemini 2.0 Flash, Claude 3.7 Sonnet, and OpenBioLLM. Accuracy was defined as the proportion of correctly answered questions in each discipline. A detailed breakdown of accuracy for each LLM across different disciplines is provided (Table 1). Overall, GPT-4o achieved the highest average accuracy (89.3%), followed by DeepSeek-R1 (87.0%) and Llama 3.3 (84.1%). Gemini 2.0 Flash reached 82.7% and Claude 3.7 Sonnet 81.2%, while OpenBioLLM and DeepSeek-V3 recorded the lowest scores at 78.2% and 76.5%, respectively.

When analyzed across individual disciplines, GPT-4o outperformed all other models, achieving the highest score in 14 of the 25 disciplines, predominantly within clinical areas such as pulmonology and infectious diseases. DeepSeek-R1 closely followed, leading in 11 disciplines, with particularly strong results in population health domains like biostatistics and public health. While Claude 3.7 Sonnet, Llama 3.3 and Gemini 2.0 Flash showed the highest accuracy in a limited number of, neither OpenBioLLM nor DeepSeek-V3 ranked highest in any of the assessed disciplines (Figure 1). Overall, there was a statistically

Table 1: Overall accuracy of each model and its performance across medical disciplines.

Medical Specialties	LLM performance, accuracy ratio (%)						
	Claude 3.7 Sonnet	DeepSeek-R1	DeepSeek-V3	Gemini 2.0 flash	GPT-4o	Llama 3.3	OpenBio
Overall							
All questions	**81.2%**	**87.0%**	**76.5%**	**82.7%**	**89.3%**	**84.1%**	**78.2%**
Allergy and immunology	77.5%	87.5%	77.5%	80.0%	82.5%	82.5%	70.0%
Anatomy and embryology	90.0%	90.0%	87.5%	87.5%	90.0%	82.5%	80.0%
Behavioral science	92.5%	90.0%	85.0%	92.5%	90.0%	85.0%	90.0%
Biochemistry	72.5%	87.5%	70.0%	75.0%	85.0%	82.5%	80.0%
Biostatistics and epidemiology	85.0%	90.0%	80.0%	77.5%	77.5%	85.0%	80.0%
Cardiology	55.0%	75.0%	52.5%	65.0%	82.5%	67.5%	75.0%
Endocrinology	85.0%	82.5%	72.5%	82.5%	90.0%	87.5%	65.0%
Gastroenterology	80.0%	90.0%	77.5%	85.0%	97.5%	90.0%	82.5%
Genetics	75.0%	75.0%	65.0%	80.0%	92.5%	80.0%	65.0%
Hematology	82.5%	92.5%	77.5%	85.0%	90.0%	85.0%	90.0%
Histology and molecular biology	82.5%	90.0%	72.5%	80.0%	90.0%	87.5%	80.0%
Infectious diseases	95.0%	90.0%	92.5%	90.0%	97.5%	85.0%	90.0%
Legal medicine and ethics	90.0%	90.0%	77.5%	80.0%	82.5%	82.5%	82.5%
Microbiology	82.5%	87.5%	77.5%	87.5%	95.0%	80.0%	70.0%
Nephrology	72.5%	90.0%	72.5%	80.0%	87.5%	87.5%	70.0%
Neurology	77.5%	77.5%	77.5%	85.0%	92.5%	80.0%	80.0%
Obstetrics and gynecology	92.5%	82.5%	75.0%	92.5%	92.5%	87.5%	70.0%
Pathology	82.5%	87.5%	70.0%	85.0%	95.0%	85.0%	75.0%
Pediatrics	85.0%	90.0%	77.5%	77.5%	87.5%	90.0%	82.5%
Physiology	77.5%	80.0%	70.0%	72.5%	80.0%	77.5%	77.5%
Psychiatry	90.0%	97.5%	85.0%	87.5%	100.0%	90.0%	87.5%
Public health	77.5%	87.5%	75.0%	77.5%	85.0%	70.0%	75.0%
Pulmonology	72.5%	87.5%	85.0%	85.0%	92.5%	85.0%	87.5%
Rheumatology	72.5%	87.5%	77.5%	87.5%	87.5%	90.0%	67.5%
Surgery	85.0%	90.0%	82.5%	90.0%	90.0%	97.5%	82.5%

LLM: Large language model, GPT: Generative pre-trained transformer

significant difference in accuracy among the seven LLMs (χ^2 test, p<0.001). Pairwise comparisons revealed that GPT-4o achieved significantly higher accuracy than DeepSeek-V3, OpenBioLLM, Claude, and Gemini 2.0 (p<0.001 for all), establishing it as the top-performing model. DeepSeek-R1 also significantly outperformed both DeepSeek-V3 and OpenBioLLM (p<0.001), demonstrating consistent high performance. Llama 3.3 scored significantly higher than DeepSeek-V3 (p<0.05). No statistically significant differences were observed between GPT-4o and DeepSeek-R1, or among Claude, Gemini 2.0, and other non-leading models (Supplementary Material S2).

Discipline-Level Performance

Infectious diseases (n=6, 91.4%), psychiatry (n=4, 91.1%), and behavioral science (n=4, 89.3%) were the disciplines in which models achieved the highest average accuracies. Conversely, the lowest-performing disciplines were cardiology (n=6, 67.5%), physiology (n=5, 76.4%), biochemistry (n=5, 78.9%), and genetics (n=4, 76.1%) (Figures 2 and 3).

To assess whether LLMs performance varied between clinical and basic sciences, the 25 medical specialties were categorized into two groups: 12 basic science disciplines and 13 clinical science disciplines. Clinical disciplines such as infectious diseases and surgery generally achieved higher scores than basic science disciplines like biochemistry, genetics, and physiology; however, this difference was not statistically significant (p=0.055), and no LLM's performance differed significantly between the two groups.

Within-Model Across Discipline Performance

Statistically significant differences in performance across medical disciplines were observed in all 7 LLM. For Claude 3.7 Sonnet, performance in cardiology was significantly lower than in disciplines such as anatomy, psychiatry, and infectious diseases (p<0.05). DeepSeek-R1 performed better in psychiatry compared to several other disciplines. DeepSeek-V3 and Gemini 2.0 both showed reduced accuracy in cardiology relative to areas like infectious diseases and surgery (p<0.05). GPT-4o scored higher in psychiatry and infectious diseases than in biostatistics and epidemiology, and physiology. Llama 3.3 performed better in surgery and psychiatry than in cardiology and public health. OpenBioLLM showed higher accuracy in behavioral science and hematology than in genetics and endocrinology (p<0.05).

Within-Discipline Across Model Performance

GPT-4o consistently outperformed other models in cardiology, gastroenterology, genetics, microbiology, pathology, and psychiatry (p<0.05). In endocrinology, both GPT-4o (p=0.014) and Llama 3.3 (p=0.034) performed better than OpenBioLLM. OpenBioLLM also showed lower performance in nephrology and obstetrics and gynecology compared to multiple models. Additionally, Claude Sonnet 3.7 and Gemini 2.0 were significantly outperformed by GPT-4o in select disciplines.

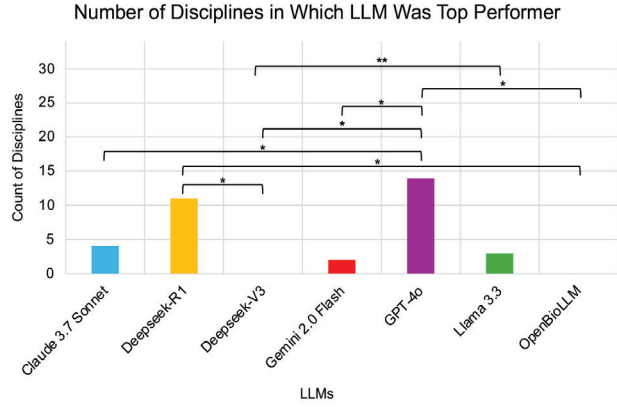


Figure 1: Number of medical disciplines in which each LLM was the top performer. This figure summarizes the distribution of first-place rankings across 25 medical disciplines. A top performer is defined as the model achieving the highest accuracy in each respective discipline. *indicates statistically significant difference at p<0.001; **Indicates statistically significant difference at p<0.05.
LLMs: Large language models, GPT: Generative pre-trained transformer

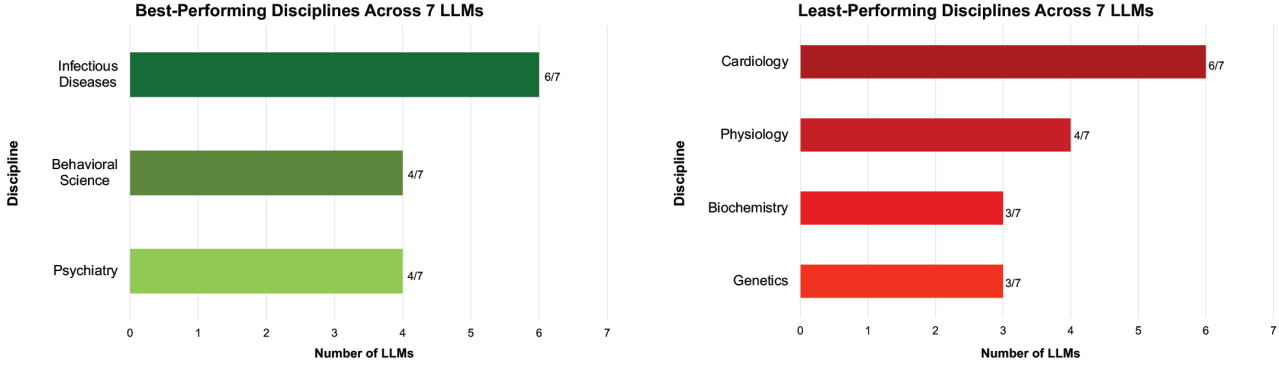


Figure 2: Best- and worst-performing medical disciplines across seven LLMs. Bars show the number of models that achieved the highest accuracy in each discipline (left) or the lowest accuracy (right), based on evaluations across 25 medical disciplines. Numbers at the end of each bar show how many models (out of 7) achieved that performance.
LLMs: Large language models

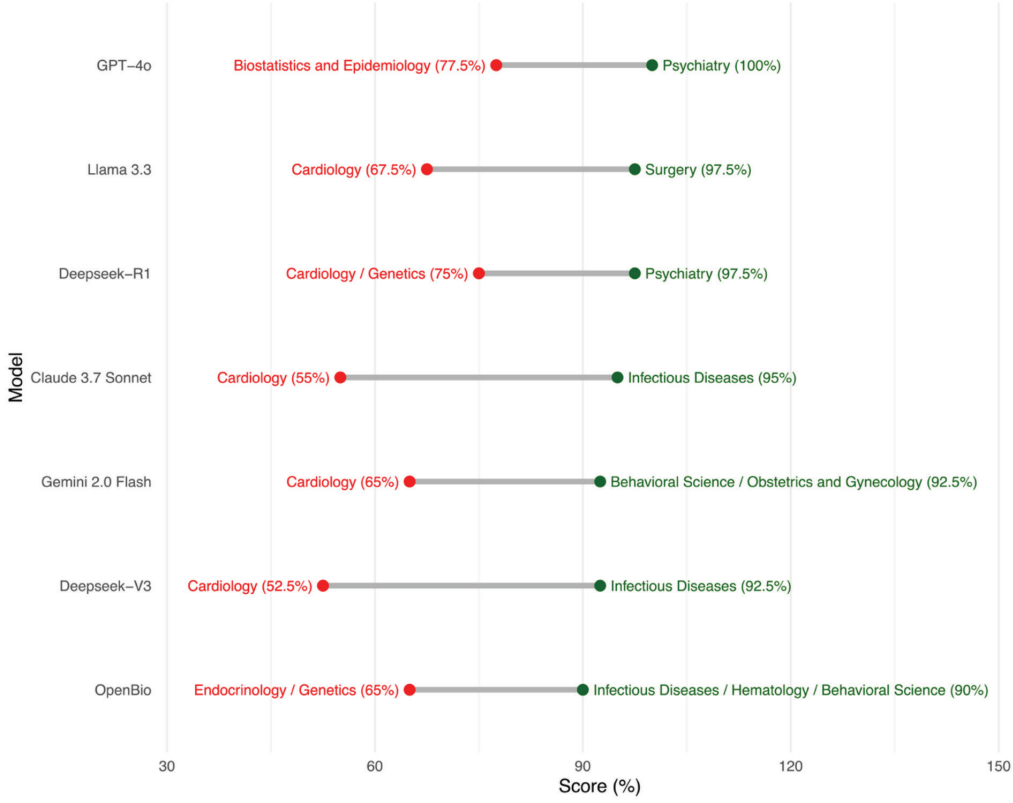


Figure 3: Best and worst performing medical disciplines for each LLM on USMLE-style questions. This dumbbell plot illustrates the highest- and lowest-performing medical disciplines for each LLM based on accuracy. Red dots indicate the lowest-performing disciplines and green dots indicate the highest-performing ones, with corresponding accuracy percentages shown in parentheses. This figure underscores the variability in domain-specific strengths and weaknesses among LLMs.
USMLE: United States Medical Licensing Examination, LLM: Large language model, GPT: Generative pre-trained transformer

When comparing across all specialties, the least variation in performance was observed in behavioral science (range 85.0% - 92.5%), whereas the greatest variation was noted in Cardiology (range 52.5% - 82.5%), highlighting disciplines where LLMs demonstrated stable versus highly divergent accuracy (Supplementary Material S3).

DISCUSSION

This study provides a comprehensive assessment of seven LLMs on 1000 USMLE-style questions from 25 medical disciplines. Among the evaluated models, GPT-4o and DeepSeek-R1 demonstrated comparable overall accuracy (89.3% and 87%, respectively), significantly outperforming DeepSeek-V3 (76.5%), OpenBioLLM (78.2%), Claude 3.7 Sonnet (81.2%), and Gemini 2.0 Flash (82.7%) ($p<0.001$). GPT's consistent success across more than half of the disciplines, particularly in clinical fields such as surgery and infectious diseases, suggests strong capabilities in both factual knowledge and applied clinical reasoning. Our findings confirm and extend prior work showing that GPT-4-based models consistently achieve high performance on medical knowledge tasks (13), underlining their potential utility in medical education and supporting earlier calls to strategically integrate high-performing LLMs into curricula (13). On the other hand, DeepSeek-R1 performed better in population health-oriented domains such as biostatistics and public health. While previous research has shown medical reasoning abilities of DeepSeek-R1, it exhibits limitations in more complex clinical scenarios (7). In contrast, OpenBioLLM and DeepSeek-V3 performed the worst, failing to lead in any single discipline. Although OpenBioLLM is specifically trained on biomedical content, its lower performance suggests that focusing only on medical material does not guarantee better overall performance in comprehensive medical exams like the USMLE.

A key finding from this study is the variation in LLM performance not only between models but also across different medical disciplines. On average, the highest-scoring areas were infectious diseases (91.4%), psychiatry (91.1%), and behavioral science (89.3%), while the lowest scores were observed in cardiology (67.5%), genetics (76.1%), and physiology (76.4%). These results suggest that certain areas of medicine are more compatible with current LLM capabilities, while others remain challenging across all models. The consistently poor performance across models in cardiology is particularly noteworthy, as this field often involves

complex cases and multiple health issues that require nuanced clinical reasoning, an area where LLMs commonly struggle (3). Our findings align with earlier studies showing that while LLMs like ChatGPT handle simple medical questions well, their performance drops with more complex clinical decision-making or specialized knowledge, sometimes producing incorrect or misleading answers (14). This may explain the lower accuracy seen in challenging areas like cardiology and genetics, where deeper reasoning is required.

When the 25 disciplines were grouped into basic sciences (e.g., biochemistry, pathology, physiology) and clinical sciences (e.g., pediatrics, surgery, infectious diseases), clinical subjects tended to score slightly higher. However, the overall difference was not statistically significant and no LLM in the study demonstrated a statistically significant difference in its own performance between basic and clinical sciences.

A strength of this study is the large and diverse question set, which systematically covers 25 medical disciplines and enables detailed comparisons across multiple models. Previous studies have compared only two or three LLMs on general question sets without focusing on discipline-specific performance. In addition, we evaluated two versions of the same LLM, allowing assessment of whether newer iterations demonstrated improved performance.

From an educational perspective, high-performing LLMs such as GPT-4o and DeepSeek-R1 could serve as useful assistants to medical training, particularly for reinforcing factual knowledge and supporting clinical reasoning in disciplines where their accuracy is consistently high. Future research should focus on expanding the analysis of USMLE-style questions by including imaging and multimedia content and covering a wide variety of clinical scenarios. This would provide a more comprehensive assessment of LLM capabilities and their ability to handle diverse, real-world clinical cases tested in the USMLE. Previous research indicates that it is important to identify which models perform better in specific contexts to enhance their practical applications, such as in diagnosis, treatment, and patient education (15). Additionally, future research is essential to improve and broaden these applications.

Study Limitations

This study contains several limitations. First, these questions are not actual USMLE exam questions, they are USMLE-style. All questions were sourced from AMBOSS, a widely used but proprietary platform. Thus, the discipline-level success rates reflect AMBOSS's specific question style and difficulty, which may limit applicability to actual exams. Future studies should use multiple question banks to improve generalizability. Second, it is important to note that no questions containing images, charts, or tables were included, in order to maintain consistency in comparison. While DeepSeek-R1 does not support image-based tasks, GPT-4o is capable of interpreting images. Lastly, as LLMs and their training data advance rapidly, the results of this work may not generalize to future iterations of these models.

CONCLUSION

In conclusion, while models like GPT-4o and DeepSeek-R1 demonstrated strong overall performance, all models showed notable variability depending on the medical discipline. While the potential of language models is considerable, it is important to interpret these findings carefully. Their limitations and risk of incorrect answers highlight the need for careful validation and further improvement before use in real healthcare or educational settings. Of note, while LLMs performed relatively well, it is important to recognize that becoming a physician involves far more than simply answering licensing exam questions correctly.

Ethics

Ethics Committee Approval: *This study did not require research ethics approval as it did not involve human subjects.*

Acknowledgments

The authors extend their gratitude to AMBOSS for providing the multiple-choice questions used in this study, which made this research possible.

Conflict of Interest: *The authors declared no conflict of interest.*

Author Contributions: *Surgical and Medical Practices: Z.S.Ö., B.B.K., E.A., Concept: Z.S.Ö., B.B.K., E.A., Design: Z.S.Ö., B.B.K., E.A., Data Collection or Processing: Z.S.Ö., B.B.K., Analysis and/or Interpretation: Z.S.Ö., E.A., Literature Search: Z.S.Ö., B.B.K., Writing: Z.S.Ö., B.B.K., E.A.*

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

REFERENCES

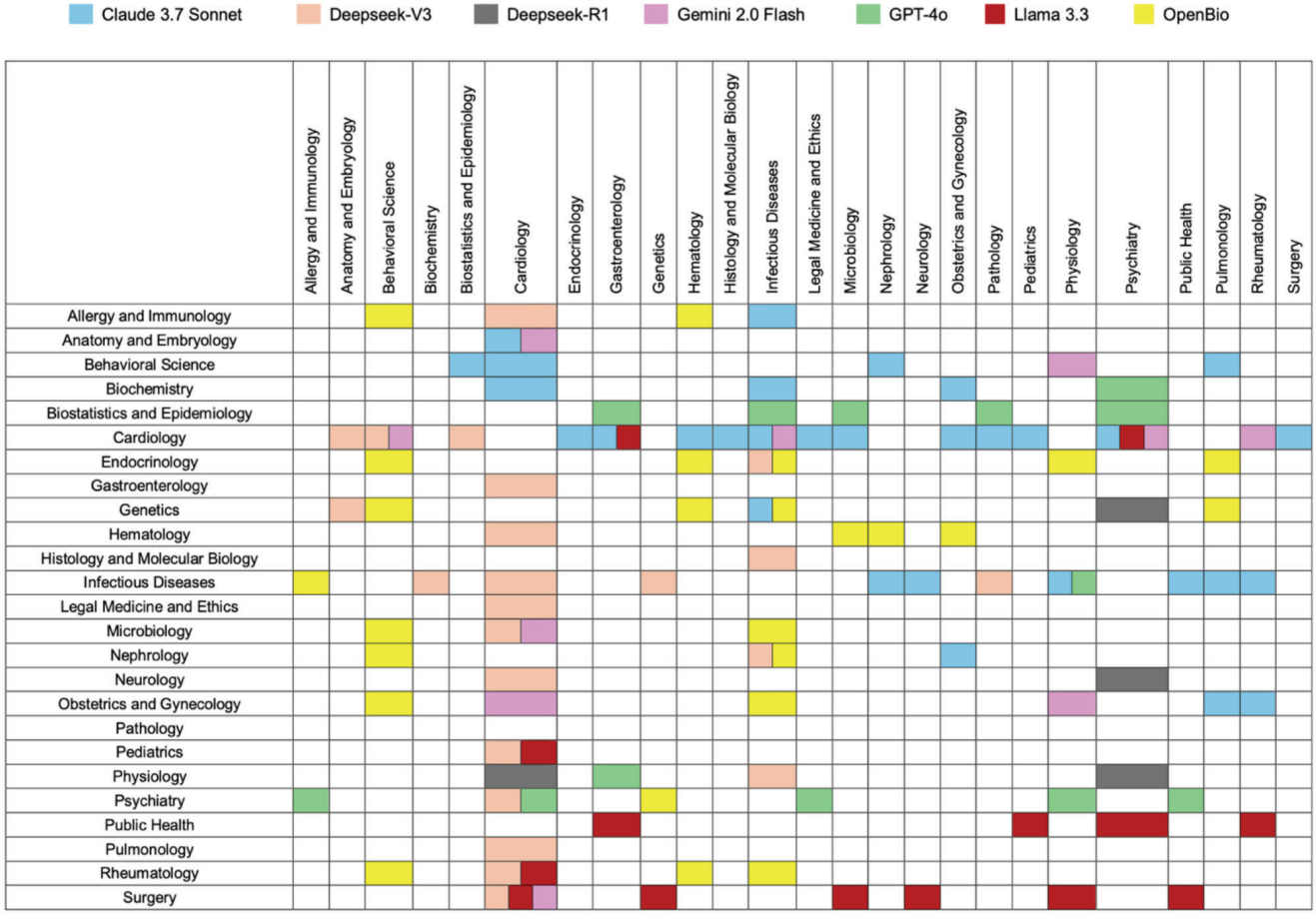
1. Gilson A, Safranek CW, Huang T et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. [Crossref]
2. Liu PR, Lu L, Zhang JY et al. Application of artificial intelligence in medicine: an overview. Curr Med Sci. 2021;41(6):1105-15. [Crossref]
3. Brin D, Sorin V, Vaid A et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13(1):16492. [Crossref]
4. Nori H, King N, McKinney SM et al. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023. [Crossref]
5. Lawrence ECN, Dine CJ, Kogan JR. Preclerkship medical students' use of third-party learning resources. JAMA Netw Open. 2023;6(12):e2345971. [Crossref]
6. Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [Crossref]
7. Moell B, Aronsson FS, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. arXiv. 2025. [Crossref]
8. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF et al. Performance of ChatGPT on the peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ. 2023;9:e48039. [Crossref]
9. Yaneva V, Baldwin P, Jurich DP et al. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med. 2024;99(2):192-7. [Crossref]
10. AMBOSS GmbH, n.d. Qbank. [online] New York: AMBOSS GmbH. Available at: https://www.amboss.com/us?_gl=1*1y682k3*_up*MQ..*_gs*MQ..&gclid=Cj0KCQjwjL3HBhCgARIsAPUg7a6aNF1mjSUBASlvBcFlgQCd-YJuCpHmk_1Y4YRMpuoFxSyRowbu0GoaArafEALw_wcB&gbraid=0AAAADQY-JPDsbHwK33oARm2lcr-mrHaH [Accessed 2 April 2025]. [Crossref]
11. Blackwell RE, Barry J, Cohn AG. Towards reproducible LLM evaluation: quantifying uncertainty in LLM benchmark scores. arXiv. 2024. [Crossref]
12. Renze M. The effect of sampling temperature on problem solving in large language models. ACL. 2024;7346-56. [Crossref]

13. Bicknell BT, Butler D, Whalen S et al. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. 2024;10:e63430. [Crossref]
14. Li S. Exploring the clinical capabilities and limitations of ChatGPT: a cautionary tale for medical applications. Int J Surg. 2023;109(9):2865-7. [Crossref]
15. Temsah MH, Jamal A, Alhasan K et al. OpenAI o1-preview vs. ChatGPT in healthcare: a new frontier in medical ai reasoning. Cureus. 2024;16(10):e70640. [Crossref]

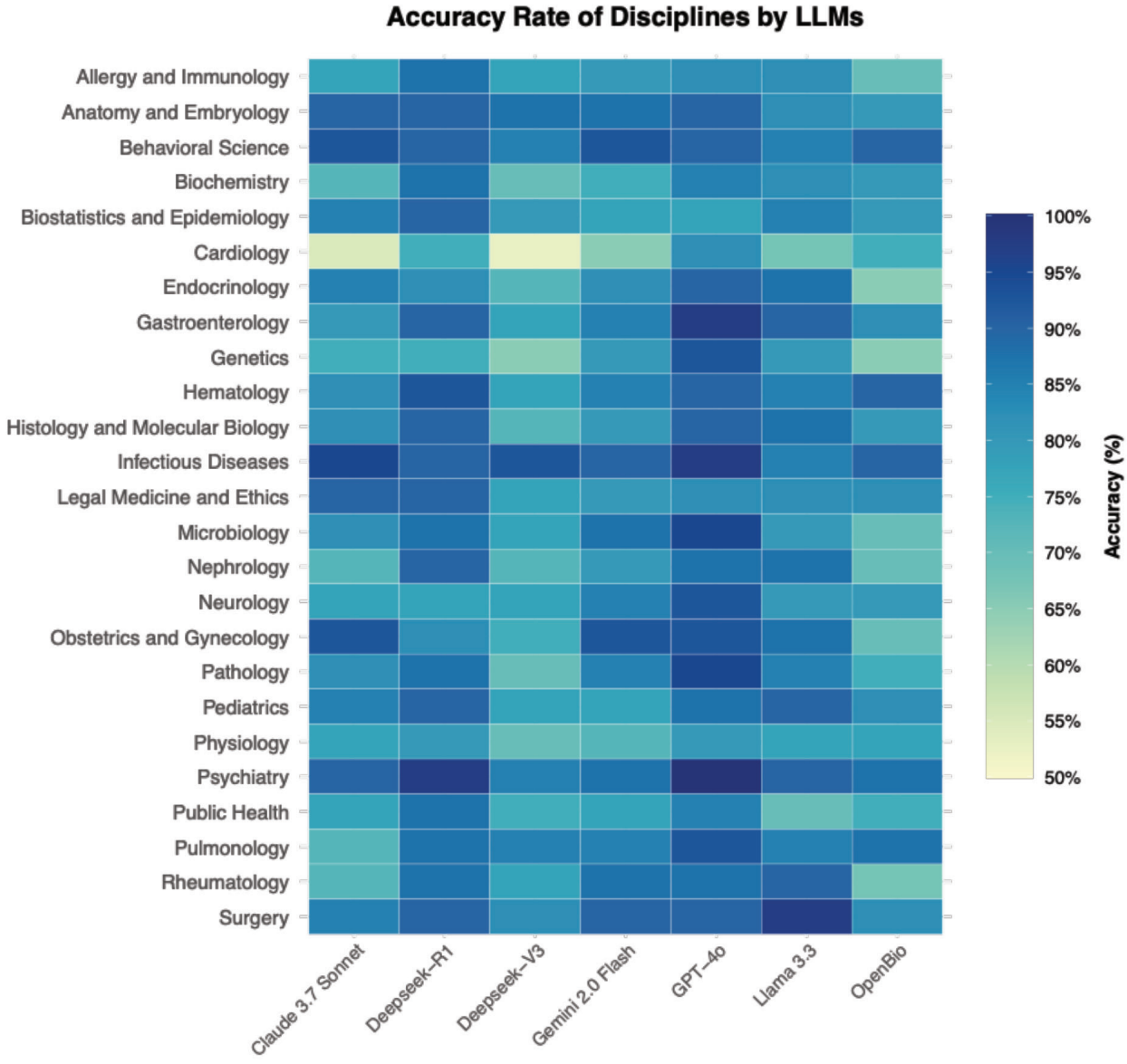
Supplementary Material S1: LLM configuration summary.

Model name	Version / identifier	API provider	Temperature	Seed support	Access date
GPT-4o	GPT-4o-2024-08-06	OpenAI API	0.0	✅ Yes	13 March 2025
Claude 3.7 sonnet	Claude-3-7-sonnet-20250219	Anthropic API	0.0	❌ No	13 March 2025
Gemini 2.0 flash	Gemini-2.0-flash (Feb 2025)	Google AI studio	0.0	✅ Yes	15 March 2025
LLaMA 3.3 70B	Meta-llama/Llama-3.3-70B-Instruct	Groq API	0.0	✅ Yes	14 March 2025
OpenBioLLM 70B	Aaditya/Llama3-OpenBioLLM-70B	Nebius API	0.0	✅ Yes	19 March 2025
DeepSeek V3	DeepSeek-ai/DeepSeek-V3	Nebius API	0.0	✅ Yes	19 March 2025
DeepSeek R1	DeepSeek-ai/DeepSeek-R1	Nebius API	0.0	✅ Yes	19 March 2025

This table provides an overview of the configuration details for each large language model evaluated in the study, including model version identifiers, API access sources, temperature settings, seed support status, and date of access. These parameters were standardized as much as possible to ensure comparability across models.
LLM: Large language model, API: Application programming interface, AI: Artificial intelligence, GPT: Generative pre-trained transformer



Supplementary Material S2: Pairwise comparisons between medical disciplines within each LLM. This cross-table shows statistically significant differences in performance between pairs of 25 medical disciplines across seven LLMs. Colored boxes indicate statistically significant differences in performance between disciplines for the corresponding LLM, with each LLM assigned a unique color (legend above the table). The absence of a colored box indicates no significant difference for any LLM. This cross-table highlights variation in discipline-specific performance across different LLMs.
LLM: Large language model, GPT: Generative pre-trained transformer



Supplementary Material S3: Heatmap showing the performance of each LLM across different medical disciplines. This heatmap illustrates the relative accuracy of LLM across 25 medical disciplines, based on responses to 1,000 USMLE-style multiple-choice questions. Each row represents a medical discipline, each column represents a LLM, and each box represents accuracy of that discipline in a particular LLM. Color intensity corresponds to performance, with darker shades indicating higher accuracy and lighter shades indicating lower accuracy (see color scale on the right).

USMLE: United States Medical Licensing Examination, LLM: Large language model, GPT: Generative pre-trained transformer