İç Hastalıkları Dergisi
Turkish Archives of Internal Medicine

RESEARCH ARTICLE/ARAŞTIRMA MAKALESİ

İç Hastalıkları Dergisi 2025;26(3-4):47-55 • **Doi:** 10.5578/ichastaliklari.20250324

The Assessment of Chatbots Performance in Internal Medicine Education

İç Hastalıkları Eğitiminde Chatbotların Performansının Değerlendirilmesi

Nur Düzen Oflas[1](İD), **Hamit Hakan Alp**[2](İD)

[1] Department of Internal Medicine, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye

[2] Department of Clinical Biochemistry, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye

Cite this article as: Düzen Oflas N, Alp HH. The assessment of chatbots performance in internal medicine education. İç Hastalıkları Dergisi 2025;26(3-4):47-55.

ABSTRACT

Introduction: *This study aimed to evaluate the performance of five current large language model–based chatbots in Internal Medicine specialty education by analyzing their accuracy on standardized examination questions.*

Materials and Methods: *This cross-sectional methodological accuracy study compared the performance of ChatGPT-5 Thinking, DeepSeek, Gemini, ChatGPT-4o, and Grok on 40 multiple-choice questions from the Turkish Internal Medicine Subspecialty Examination. All questions were presented in English to isolate medical knowledge from language proficiency. Each question was asked in a separate session, with no contextual memory carried forward. The final selected answer was compared to the official reference key. Accuracy rates and pairwise statistical comparisons were performed using the McNemar test with Bonferroni correction.*

Results: *Accuracy rates were 97.5% for ChatGPT-5 Thinking, 90.0% for DeepSeek, 87.5% for Gemini, 87.5% for ChatGPT-4o, and 82.5% for Grok. Pairwise comparisons showed limited statistically significant differences, indicating convergence in factual knowledge retrieval among recent models. However, occasional confidently presented incorrect responses were observed, reflecting risks associated with hallucination and over-generalization.*

Conclusion: *Large language model–based chatbots demonstrate high accuracy on Internal Medicine examination questions and may serve as supportive learning tools in specialty education. However, due to the potential for misleading or overconfident incorrect reasoning, these tools should currently complement-not replace-expert-guided instruction. Structured supervision and critical appraisal training are essential for safe educational integration.*

Key Words: *Chatbots, large language models, artificial intelligence, medical education, internal medicine*

Address for Correspondence/Yazışma Adresi

Nur Düzen Oflas

Department of Internal Medicine, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye

e-mail: dr.nurdzn@hotmail.com

Received: 05.11.2025
Accepted: 15.12.2025
Available Online Date: 29.12.2025

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

ÖZ

Giriş: *Bu çalışmada, güncel büyük dil modeli tabanlı chatbotların iç hastalıkları uzmanlık eğitiminde kullanılabilirliğini değerlendirmek amacıyla, standart sınav sorularındaki doğruluk performansları incelendi.*

Materyal ve Metod: *Kesitsel metodolojik doğruluk çalışması olarak tasarlanan bu araştırmada ChatGPT-5 Thinking, DeepSeek, Gemini, ChatGPT-4o ve Grok modelleri, İç Hastalıkları Yandal Uzmanlık Sınavı'ndan seçilen 40 çoktan seçmeli soru ile test edildi. Sorular, dil yeterlilik etkisini ortadan kaldırmak amacıyla İngilizce sunuldu. Her soru ayrı oturumda soruldu ve modellerin son cevapları ÖSYM'nin resmi anahtarı ile karşılaştırıldı. McNemar testi Bonferroni düzeltmesi ile kullanılarak ikili karşılaştırmalar yapıldı.*

Bulgular: *Doğruluk oranları ChatGPT-5 Thinking için %97.5, DeepSeek için %90.0, Gemini için %87.5, ChatGPT-4o için %87.5 ve Grok için %82.5 olarak bulundu. İkili analizler, modeller arasında sınırlı anlamlı farklılık olduğunu ve son nesil modellerin bilgi geri çağırma düzeylerinde yakınlaştığını gösterdi. Ancak bazı sorularda kendinden emin şekilde verilen hatalı yanıtlar gözlendi.*

Sonuç: *Chatbotlar, iç hastalıkları uzmanlık eğitiminde destekleyici öğrenme aracı olarak potansiyel taşımaktadır. Bununla birlikte, yanlış ve güvenli olmayan çıkarımlar yapabilme riskleri nedeniyle, şu aşamada uzman eğitimi ve gözetimi yerine geçmemelidir. Güvenli entegrasyon için denetimli kullanım ve eleştirel değerlendirme becerilerinin geliştirilmesi gereklidir.*

Anahtar Kelimeler: *Chatbotlar, büyük dil modelleri, yapay zeka, tıp eğitimi, iç hastalıkları eğitimi*

INTRODUCTION

Generative artificial intelligence (GENERATIVE AI) and large language models (LLMs) that underpin this technology (Generative Artificial Intelligence, GENERATIVE AI) have the potential to trigger a paradigm shift in medicine and healthcare (1). These models find applications in a wide range of areas from the development of clinical decision support systems to the personalisation of patient education materials and to the interpretation of complex medical data and the automation of administrative workloads (2-4). This rapid transformation in the medical field has led prominent medical publications such as the New England Journal of Medicine (NEJM) to establish specialised journals such as NEJM AI, confirming that AI is no longer a fad but an integral part of medical decision-making processes (5).

Despite this tremendous potential, the integration of LLMs into clinical practice faces significant obstacles, including accuracy, reliability, and ethical imperatives. These models' tendency to generate false information—known as "hallucination"—that sounds plausible but lacks a factual basis poses a direct risk to patient safety (6). Furthermore, issues such as the models' reflection of implicit biases in the training data, lack of interpretability, and data privacy are critical research areas that must be meticulously addressed before these technologies can be responsibly deployed (7). Therefore, continuous benchmarking of models' medical knowledge and clinical reasoning abilities using objective, standardised, and reproducible methods has become a necessity (8).

In response to this assessment need, standardised medical competency exams have been adopted in the scientific literature as the de facto gold standard for measuring the knowledge level of LLMs. In a pioneering work in this area, early versions of ChatGPT achieved a significant milestone by demonstrating "passing" or "near-passing" performance on the United States Medical Licensing Exam (USMLE) (9). Subsequently, rapid advances in model architectures have dramatically improved performance. For example, the GPT-4 model has been reported to achieve high accuracy rates ranging from 80% to 88% on different steps of the USMLE (10). Med-PaLM 2, developed by Google and optimised specifically for the medical field, was one of the first models to exceed the "expert level" with an accuracy rate of 86% (11).

Current literature demonstrates that this performance bar is continually rising. Next-generation models such as Med-Gemini have achieved 91% accuracy on the USMLE (MedQA) dataset, and GPT-4o has achieved 90% accuracy, surpassing human expert performance (4). Similarly, models such as DeepSeek have demonstrated exceptional accuracy rates of 98% on pediatric board exams, 92% on the USMLE, demonstrating that LLMs are rapidly approaching superhuman levels in their ability to recall and apply standard medical knowledge (12,13).

However, the vast majority of existing LLM evaluation studies are USMLE-centric, focusing primarily on English-language data and the US medical education context (14). This creates a significant knowledge gap regarding the performance of models in different

languages across different medical education systems and on national qualifying exams with varying question structures. Recently, studies examining non-English language exams, such as the Japanese National Medical Qualification Examination, the Chinese National Medical Qualification Examination, and the Thai Medical Qualification Examination, have begun to fill this gap (14-16). These studies have revealed that model performance can vary depending on language, question difficulty, and speciality. In this context, the Subspecialty Exam for Internal Medicine Specialists (YDUS; YanDal Uzmanlık Sınavı), the basic qualifying exam required of medical doctors in Türkiye to receive speciality training, offers a critical local context for assessing LLM performance (17). The limited number of previous studies on the YDUS has demonstrated the potential of these models. For example, a study by Kılıç (2023) reported that GPT-4 outperformed GPT-3.5 (40.17%) and even the average of physicians taking the exam (38.14%) on YDUS questions (70.56%) (18).

Our aim in this study is to fill this gap in the literature by evaluating the performance of the five most current major language models (Gemini, ChatGPT-5 Thinking, DeepSeek, ChatGPT-4o, and Grok) on internal medicine questions of the YDUS through a methodological accuracy study. Our study differs from previous YDUS studies in one important respect: our methodology uses a standardised approach where all questions are presented to the models in English. This methodological choice allows us to directly test the extent to which the models have acquired the basic medical knowledge and clinical reasoning required by the YDUS questions rather than their Turkish language proficiency. This study is one of the first to examine the performance of newer-generation models (such as Grok and DeepSeek) in the YDUS context, conducting this assessment in English, independent of the language factor (Turkish).

MATERIALS and METHODS

This study is a cross-sectional, methodological accuracy evaluation comparing the response accuracy of five major language models (Gemini, ChatGPT-5 Thinking, DeepSeek, ChatGPT-4o, and Grok) on multiple-choice questions on the YDUS. The primary outcome measure was the question-by-question correct response rate (accuracy) of each model. All prompts and answers were in English. The exam questions were translated into English manually by the research team to ensure semantic accuracy and consistency.

Data Source and Question Selection

The question population was composed of multiple-choice questions from the YDUS exams publicly disclosed by Assessment, Selection, and Placement Center. Due to copyright restrictions, the questions were not republished in the study; they were used solely for the purpose of evaluating the question body and options. A total of 40 questions were selected to represent different years. Inclusion Criteria: The questions were included in a multiple-choice format (A–E) and were verifiable with an answer key on the Assessment, Selection, and Placement Center official website.

Items with errata/cancellation notices were excluded from the study. Questions with spelling errors were also reviewed by two independent researchers, and any questions containing spelling errors were excluded from the study. Questions were selected using a probability stratified sampling principle based on subject area and year distribution. The final list was obtained by independent review and consensus of two researchers.

Gold Standard (Reference)

The gold standard answers are the official Assessment, Selection, and Placement Center answer keys for the relevant exams. The correct answer choices (A–E) for each item were recorded in a blinded file before the study began and were hidden from the evaluators during the application.

Compared Models and Runtime Environment

Five chatbots were evaluated: Gemini, ChatGPT-5 Thinking, DeepSeek, ChatGPT-4o, and Grok. Each model was run through its provider's official interface: Access and date range: [October 3 -November 30, 2025]. Language: All prompts and answers were in English. Session isolation: Each question was asked in a new speaking session, and the context of the previous conversation was not carried over. Parameters: For user-configurable models, temperature=0.0, top_p=1.0, and max. output length are standardized to ≥[2048] tokens. Models that did not offer parameter control were used with default settings, and this was reported. Tools/Additional Plugins: External tools such as scanning, web search, and code execution were disabled (to the extent possible). Model storage/personalization features were disabled or a clean session was used. The questions were translated into English by the researchers to ensure linguistic consistency.

Prompt Standardization

A single round of interaction with the model was implemented for each question. Prompt template:

"Answer the following YDUS multiple-choice question using only a single letter (A, B, C, D, or E) on the last line. Do not provide a rationale.

Question: [question body]

Options: A) ... B) ... C) ... D) ... E) ..."

Questions were transcribed verbatim; Items requiring figures, tables, or visuals were excluded (if any, they were eliminated according to the pre-exclusion criteria). No second prompt, hint, or feedback was provided. If the model produced more than one letter, the last (bottom) letter was considered the "final guess." Outputs that contained no letters, were ambiguous, or refused to answer were considered incorrect (Fail).

Response Capture and Scoring

For each model-question pair, the raw output, date/time stamp, session ID, and system/configuration logs (if applicable) were recorded. The final guess was extracted using a regular expression (A–E) and compared to the gold standard.

A binary score was assigned: Correct = Pass, Incorrect/Empty/Multiple = Fail.

Data entry was independently checked by two researchers to prevent double-entry errors; discrepancies were resolved by a third researcher.

Blinding and Randomization

Model operators were blinded to the gold standard response for each item. The order of question presentation was randomized for each model (repeatable using Python random.seed=[...]). To prevent a fixed pattern between models, the order of assessments across daily sessions was balanced using a Latin square approach.

RESULTS

In this study, 40 multiple-choice questions from different years of the YDUS exam were evaluated by posing the same set of questions to each model in separate sessions, and responses were recorded in a binary (Pass/Fail) format. Due to the paired design, pairwise comparisons were conducted using the McNemar test, and a Bonferroni correction was applied for multiple trials (α_Bonferroni=0.0083) (Table 1). Adjusted p-values are shown in the upper triangle of Table 1, and statistically significant differences are indicated as "Yes" in the lower triangle. Raw accuracy counts, coded 0 = Pass and 1 = Fail, were distributed across models as follows: ChatGPT-5 Thinking 39 correct / 1 incorrect (97.5%), DeepSeek 36 / 4 (90%), Gemini 35 / 5 (87.5%), ChatGPT-4o 35 / 5 (87.5%), and Grok 33 / 7 (82.5%) (Figure 1). Pairwise comparisons revealed that, after Bonferroni correction, significant differences were found for Gemini and ChatGPT-4o and ChatGPT-5 and ChatGPT-4o pairings; for other pairings (comparisons with ChatGPT-5 and Gemini, ChatGPT-5 and DeepSeek, DeepSeek and Gemini, DeepSeek and ChatGPT-4o, and Grok), significance was not reached

Table 1. Pairwise comparisons of PASS rates on YDUS questions of five AI chatbot models (DeepSeek, Gemini, ChatGPT-4o, ChatGPT-5 thinking, and Grok) (McNemar tests; Bonferroni corrected p-values, α-Bonferroni= 0.0083)

	ChatGPT-5 T	DeepSeek	Gemini	ChatGPT-4o	Grok
ChatGPT-5 T	1	0.375	0.125	0.219	0.07
DeepSeek	0.375	1	1.00	1.00	0.250
Gemini	0.125	0.101	1	1.00	0.625
ChatGPT-4o	0.219	0.025	0.000	1	0.625
Grok	0.07	0.250	0.625	0.625	1
	ChatGPT-5 T	**DeepSeek**	**Gemini**	**ChatGPT-4o**	
ChatGPT-5 T	No	No	No	No	No
DeepSeek	No	No	No	No	No
Gemini	No	No	No	No	No
ChatGPT-4o	No	No	Yes	No	No
Grok	No	No	Yes	No	No

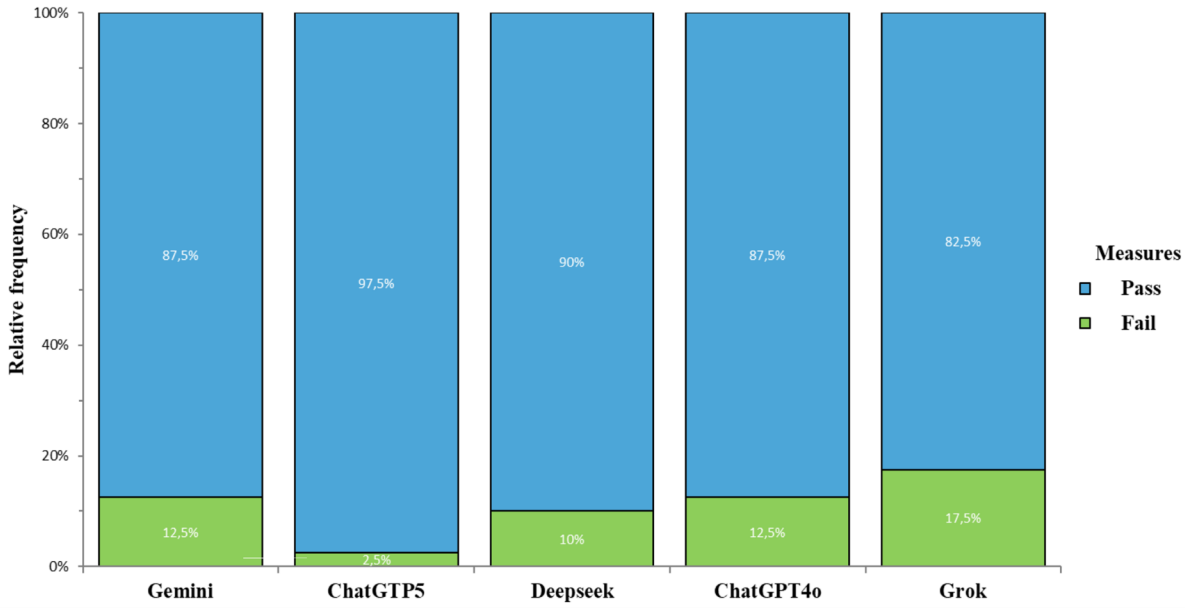

Figure 1. Distribution of Pass and Fail outcomes among the evaluated language models, expressed as relative frequencies (%).

(adjusted p≥ 0.0083). Taken together, these findings suggest that the performance heterogeneity across the examined question set is primarily due to ChatGPT-4o's relative weakness in question-level comparisons; Gemini and ChatGPT-5 Thinking are similarly accurate, DeepSeek performs close to this group, and Grok remains relatively less accurate.

Table 2 reveals a dominant pattern that holds for all model pairs: the frequencies in the "Pass-Pass" cell overwhelmingly outnumber all other cells. For example, in the ChatGPT-5 vs. DeepSeek comparison, the models collectively responded "Pass" on 35 out of 40 items (87.5%). Similarly, the ChatGPT-5 vs. ChatGPT-4o pair showed "Pass-Pass" agreement on 34/40 items (85%), and the DeepSeek vs. Grok pair showed "Pass-Pass" agreement on 33/40 items (82.5%) (Table 2). These high "Pass-Pass" frequencies initially indicate a high rate of raw agreement. Raw agreement is calculated as the ratio of the diagonal cells (Pass-Pass + Fail-Fail) to the total number of items. For example, the raw agreement ratio for the ChatGPT-5 – ChatGPT-4o pair is (34 + 0)/40= 0.85 (85%). For the ChatGPT-5 – DeepSeek pair, this ratio is (35 + 0)/40= 0.875 (87.5%). However, caution should be exercised in interpreting these high raw agreement ratios. The data in Table 2 indicate a high prevalence of the "Successful" category, indicating that the 40-item task set was likely "easy" for the majority of the models. Statistically, when the marginal distributions are this unbalanced (i.e., one category appears much more frequently than the other), the probability that two raters (or models) will agree on the same dominant category simply by chance also increases. If the two models had a 90% chance of responding "Pass" regardless of the questions, the expected "Pass-Pass" agreement by chance would already be 0.90 × 0.90= 0.81 (81%). In this context, the observed 85% raw agreement may not prove true agreement between the models. This potential for a "kappa paradox" creates the methodological need for a chance-corrected agreement coefficient beyond the raw agreement percentages. To assess the true consistency between the models' response patterns, a Cohen's Kappa analysis will be presented in the next section.

In order to assess response consistency across the models, eliminating chance factors, Cohen's Kappa (κ) coefficient, 95% confidence interval (CI), and p-value were calculated for each model pair using the 2x2 contingency tables presented in Table 2. The findings are summarized in Table 3. Cohen's Kappa measures the extent to which the observed agreement exceeds that expected by chance; kappa= 0 indicates chance-level agreement, and kappa= 1 indicates perfect agreement. Cohen's kappa coefficients were calculated for each model pair based on binary responses (0= Pass, 1= Fail) to the same 40 questions,

Table 2. Pairwise cross-classification (2×2) of PASS/FAIL outcomes between AI chatbots (n= 40)

		ChatGPT-5	
		Pass	**Fail**
DeepSeek	Pass	35	1
	Fail	4	0
Gemini	Pass	35	0
	Fail	4	1
ChatGPT-4o	Pass	34	1
	Fail	5	0
Grok	Pass	32	1
	Fail	7	0
		DeepSeek	
		Pass	**Fail**
Gemini	Pass	34	1
	Fail	2	3
ChatGPT-4o	Pass	34	1
	Fail	2	3
Grok	Pass	33	0
	Fail	3	4
		Gemini	
		Pass	**Fail**
ChatGPT-4o	Pass	33	2
	Fail	2	3
Grok	Pass	32	1
	Fail	3	4
		ChatGPT-4o	
		Pass	**Fail**
Grok	Pass	32	1
	Fail	3	4

and two-sided p-values with 95% confidence intervals are reported (Table 3). The highest agreement was observed for DeepSeek–Grok (κ= 0.688; 95% CI −0.574–0.812; p= 0.001), ChatGPT-4o–DeepSeek (κ= 0.630; 95% CI 0.38–0.88; p= 0.001), and ChatGPT-4o–Grok (κ= 0.610; 95% CI 0.34–0.75; p= 0.001) pairs. ChatGPT-5–DeepSeek matching also showed significant, moderate agreement (κ= 0.470; 95% CI 0.18–0.75; p= 0.001). For ChatGPT-5–Gemini, kappa was low but statistically significant (κ= 0.304; 95% CI 0.12–0.55; p= 0.007). In contrast, no significant agreement was found for ChatGPT-5–ChatGPT-4o (κ= 0.043; p= 0.702), ChatGPT-5–Grok (κ= 0.060; p= 0.700), ChatGPT-4o–Gemini (κ= 0.250; p= 0.100), Gemini–DeepSeek (κ= 0.100; p= 0.500), and Gemini–Grok (κ= 0.100; p= 0.500) pairs. Overall, agreement levels varied significantly between the pairs; they were particularly high for ChatGPT-4o–DeepSeek and ChatGPT-4o–Grok pairs but weak or insignificant comparisons between Gemini and the other models. The wide confidence interval for DeepSeek–Grok, crossing zero, indicates higher estimation uncertainty in this pair, while the p-value <0.05 statistically supports a deviation from chance agreement. Detailed kappa, confidence interval, and p-value values for all pairs are in Table 3.

DISCUSSION

In this study, we evaluated the performance of five large language models—ChatGPT-5 Thinking, DeepSeek,

Table 3. Cohen's Kappa coefficient, 95% confidence interval, and p-value between different chatbot pairs for 40 questions

Chatbot Couple	κ (Kappa)	95% CI	p-value
ChatGPT-5 – ChatGPT-4o	0.043	[-0.38 to 0.88]	0.702
ChatGPT-5 – Gemini	0.304	[0.12 to 0.55]	0.007
ChatGPT-5 – DeepSeek	0.47	[0.18 to 0.75]	0.001
ChatGPT-5 – Grok	0.06	[-0.23 to 0.35]	0.70
ChatGPT-4o – Gemini	0.25	[-0.05 to 0.55]	0.10
ChatGPT-4o – DeepSeek	0.63	[0.38 to 0.88]	0.001
ChatGPT-4o – Grok	0.610	[0.34 to 0.75]	0.001
Gemini – DeepSeek	0.10	[-0.15 to 0.35]	0.50
Gemini – Grok	0.10	[-0.20 to 0.40]	0.50
DeepSeek – Grok	0.688	[0.574 to 0.812]	0.001

Gemini, ChatGPT-4o, and Grok—on Internal Medicine questions selected from the Turkish subspecialty certification exam. Overall, the results demonstrated high accuracy across models, with ChatGPT-5 Thinking achieving the highest performance (97.5%), followed by DeepSeek (90.0%), Gemini (87.5%), ChatGPT-4o (87.5%), and Grok (82.5%). Pairwise comparisons suggested that performance differences between the models were limited, indicating that recent-generation LLMs and GENERATIVE AI appear to be converging toward similar levels of factual recall ability in standardized knowledge assessments.

However, high accuracy in exam-style questions does not necessarily equate to reliable clinical reasoning or safe application in real-world contexts. Previous research has emphasized that even when LLMs and GENERATIVE AI successfully answer structured medical questions, they may still generate incorrect, fabricated, or overconfident clinical statements—referred to as hallucinations—when confronted with ambiguous or context-dependent tasks (11,19,20). Such errors are not always obvious to learners and may be presented with strong linguistic confidence, increasing the risk of reinforcing misconceptions. This is particularly critical in medical education, where novices may lack the expertise to recognize subtle inaccuracies.

Previous studies evaluating LLMs and GENERATIVE AI performance across different national medical examinations have reported variable outcomes depending on language and exam structure. For example, Gilson et al. showed that ChatGPT could achieve near-passing performance on the USMLE, yet answer interpretation depth varied across question types (19). Studies conducted in China and Japan reported that performance may decline when questions involve nuanced clinical interpretation rather than direct recall (14,15). In Türkiye, Kılıç demonstrated that GPT-4 outperformed GPT-3.5 and the average test taker on specialization exam questions presented in Turkish (18). However, those studies did not separate medical knowledge from language proficiency. In contrast, the present study used English prompts exclusively, thereby evaluating medical conceptual understanding independent of the Turkish language, allowing a clearer assessment of domain knowledge versus language-dependent pattern recognition.

Despite their limitations, LLMs and GENERATIVE AI hold meaningful potential in medical training. They can provide rapid feedback, flexible self-paced learning, and opportunities for diagnostic reasoning practice (21,22). Yet, the possibility of confidently delivered but incorrect information presents both pedagogical and ethical concerns. Errors from LLMs and GENERATIVE AI are often systematic rather than random and may appear authoritative (23). Therefore, these systems should currently be considered supplementary educational tools not replacements for supervised instruction.

Moreover, real-world clinical practice requires cognitive integration, contextual judgment, and patient-specific decision-making, which are not fully captured in multiple-choice testing formats. LLMs and GENERATIVE AI currently lack the ability to consistently evaluate risk-benefit trade-offs, ethical implications, and individualized patient variability—skills essential in Internal Medicine.

LLMs and GENERATIVE AI may be used effectively as practice aids for specialty examination preparation. They should not currently be used as primary or unsupervised instructional sources. Learners should receive explicit training in critical appraisal of AI-generated outputs. Educators should supervise AI-assisted learning to prevent reinforcement of incorrect reasoning patterns.

CONCLUSION

This study demonstrates that recent LLMs and GENERATIVE AI perform well on Internal Medicine specialty examination questions, indicating meaningful potential as auxiliary tools in medical education. However, the presence of occasional but confidently delivered incorrect responses emphasizes the need for cautious, supervised, and structured integration. LLMs and GENERATIVE AI should complement—not replace—expert-guided learning. Continued refinement is required to reduce hallucinations, improve reasoning transparency, and ensure safe and effective educational use.

ETHICS COMMITTEE APPROVAL

This study was approved by the Van Yüzüncü Yıl University Non-Invasive Clinical Research (Decision no: 2025/01-40, Date: 04.02.2025).

CONFLICT of INTEREST

No conflict of interest declared.

FUNDING SOURCES

The author(s) received no financial support for the research, authorship, and/or publication of this article.

AUTHORSHIP CONTRIBUTIONS

Concept and Design: NDO

Analysis/Interpretation: HHA

Data Collection or Processing: NDO

Writing: NDO

Review and Correction: HHA

Final Approval: HHA

REFERENCES

1. *Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience 2024;27(5):109713. https://doi.org/10.1016/j.isci.2024.109713*
2. *Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare 2025;13(6):603. https://doi.org/10.3390/healthcare13060603*
3. *Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: A scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. https://doi.org/10.3389/fmed.2024.1477898*
4. *Ng FYC, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, et al. Artificial intelligence education: An evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023;4(10):101230. https://doi.org/10.1016/j.xcrm.2023.101230*
5. *Cooper A, Rodman A. AI and medical education-a 21st-century Pandora's box. N Engl J Med 2023;389(5):385-7. https://doi.org/10.1056/NEJMp2304993*
6. *Zhang P, Shi J, Kamel Boulos MN. Generative AI in medicine and healthcare: moving beyond the 'peak of inflated expectations'. Future Internet 2024;16(12):462. https://doi.org/10.3390/fi16120462*
7. *Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering 2025;12(6):631. https://doi.org/10.3390/bioengineering12060631*
8. *Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med 2024;177(2):210-20. https://doi.org/10.7326/M23-2772*
9. *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 Digital Health 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198*
10. *Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. Journal of Medical Internet Research 2025;27:e65146. https://doi.org/10.2196/65146*
11. *Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature Medicine 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8*
12. *Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Medical Education 2024;10(1):e63430. https://doi.org/10.2196/63430*

13. *Mansoor M, Ibrahim A, Hamide A. Performance of DeepSeek and GPT models on pediatric board preparation questions: Comparative evaluation. JMIR AI 2025;4:e76056. https://doi.org/10.2196/76056*

14. *Yao Z, Duan L, Xu S, Chi L, Sheng D. Performance of Large Language Models in the Non-English Context: Qualitative Study of Models Trained on Different Languages in Chinese Medical Examinations. JMIR Medical Informatics 2025;13(1):e69485. https://doi.org/10.2196/69485*

15. *Liu M, Okuhara T, Dai Z, Huang W, Okada H, Furukawa E, et al. Performance of advanced large language models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese medical licensing examination: A comparative study. medRxiv 2024;2024:07. https://doi.org/10.1101/2024.07.09.24310129*

16. *Saowaprut P, Wabina RS, Yang J, Siriwat L. Performance of large language models on Thailand's national medical licensing examination: A cross-sectional study. Journal of Educational Evaluation for Health Professions 2025;22:16. https://doi.org/10.3352/jeehp.2025.22.16*

17. *Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. Journal of Medical Internet Research 2024;26:e60807. https://doi.org/10.2196/60807*

18. *Kılıç ME. AI in medical education: A comparative analysis of GPT-4 and GPT-3.5 on Turkish medical specialization exam performance. medRxiv 2023:2023.07. 12.23292564. https://doi.org/10.1101/2023.07.12.23292564*

19. *Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How well does ChatGPT do on the USMLE? JMIR Med Educ 2023;9:e45312. https://doi.org/10.2196/45312*

20. *Shah N, Pfeffer M, Liang P. Holistic evaluation of large language models for medical applications. arXiv 2025;2505:2380.*

21. *Festor P, Jia Y, Gordon AC, Faisal AA, Habli I, Komorowski M. Assuring the safety of AI-based clinical decision support systems: a case study of the AI Clinician for sepsis treatment. BMJ Health & Care Informatics 2022;29(1):e100549. https://doi.org/10.1136/bmjhci-2022-100549*

22. *Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learning Environments 2024;11(1):28. https://doi.org/10.1186/s40561-024-00316-7*

23. *Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Medical Teacher 2024;46(4):446-70. https://doi.org/10.1080/0142159X.2024.2314198*