# How Large Language Models Perform on the United States Medical Licensing Examination: A Systematic Review

Dana Brin, MD[1,2]; Vera Sorin, MD[1-3]; Eli Konen, MD[1,2]; Girish Nadkarni, MD[4-5]; Benjamin S Glicksberg, MD[6]; Eyal Klang, MD[1-5]

[1]Department of Diagnostic Imaging, Chaim Sheba Medical Center, Tel Hashomer, Israel

[2]The Faculty of Medicine, Tel-Aviv University, Israel

[3]DeepVision Lab, Chaim Sheba Medical Center, Tel Hashomer, Israel

[4]Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, New York, USA

[5]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[6]Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Corresponding Author:

Dana Brin, MD

Department of Diagnostic Imaging, Chaim Sheba Medical Center

Address: Emek Haela St. 1, Ramat Gan, Israel, 52621.

Tel: +972-3-5302530, Fax: +972-3-5357315, Email: dannabrin@gmail.com

**Abstract**

**Purpose**: The United States Medical Licensing Examination (USMLE) assesses physicians' competency and passing is a requirement to practice medicine in the U.S. With the emergence of large language models (LLMs) like ChatGPT and GPT-4, understanding their performance on these exams illuminates their potential in medical education and healthcare.

**Methods**: A literature search following the 2020 PRISMA guidelines was conducted, focusing on studies using official USMLE questions and publicly available LLMs.

**Results**: Three relevant studies were found, with GPT-4 showcasing the highest accuracy rates of 80-90% on the USMLE. Open-ended prompts typically outperformed multiple-choice ones, with 5-shot prompting slightly edging out zero-shot.

**Conclusion**: LLMs, especially GPT-4, display proficiency in tackling USMLE-standard questions. While the USMLE is a structured evaluation tool, it may not fully capture the expansive capabilities and limitations of LLMs in medical scenarios. As AI integrates further into healthcare, ongoing assessments against trusted benchmarks are essential.

## Introduction

The United States Medical Licensing Examination (USMLE) is a multi-part professional exam that is mandatory to practice medicine in the United States. It is a rigorous assessment of physicians' knowledge and skills, providing a standardized measure of competence for both domestic and international medical graduates (1–4). As such, the USMLE has become a critical benchmark for medical education and is increasingly used in research as a standard for testing the capabilities of various healthcare-focused artificial intelligence tools.

Advancements in natural language processing (NLP), have led to the development of large language models (LLMs) like GPT-3 and GPT-4. These models can generate human-like text and perform various complex NLP tasks. LLMs are increasingly studied in healthcare for different applications, including aiding diagnosis, streamlining administrative tasks, and enhancing medical education (5–8). It is critical to assess the performance of these models in a standardized and rigorous manner, especially in specialized fields like medicine (6,9,10).

Given the important role of the USMLE in assessing medical competence, understanding LLMs fare on this test offers valuable insights into their clinical reasoning, potential applications, and limitations in healthcare. Thus, the aim of our study was to systematically review the literature on the performance of publicly available LLMs on official USMLE questions, analyze their clinical reasoning capabilities, and determine the impact of various prompting methodologies on outcomes.

## Methods

### Literature search

A systematic literature search was conducted for studies on LLM's performance on the USMLE. We searched for articles published up to July 2023. Pubmed and Google Scholar were used as databases. Search keywords included "USMLE", "United Stated Medical License Exams", "LLM" and "Large Language Models". We also searched the references lists of relevant studies for any additional relevant studies. **Figure 1** presents a flow diagram of the screening and inclusion process.

### Eligibility Criteria

We included studies that evaluated the performance of publicly available LLM's on official USMLE questions. We excluded papers that evaluate LLMs using unofficial sources of USMLE-like questions (e.g., MedQA) and non-English papers.

### Screening and Synthesis

This review was reported according to the 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (10).

## Results

Data were extracted from three publications that evaluated the performance of publicly available LLMs on USMLE questions (**Table 1**). The parameters evaluated in each publication are described in **Table 2**. The results are shown in **Tables 3-5**.

### Data sets

There are two official sources for USMLE questions – USMLE Sample exam, which is freely available (11), and the NMBE Self-Assessment, available for purchase at the NMBE website (12). Both include questions for Step 1, 2CK and 3.

## Large Language Models

The large language models included in this study were all developed by OpenAI (13). GPT-3 is an autoregressive model known for its ability to handle a variety of language tasks without extensive fine-tuning. GPT-3.5, a subsequent version, serves as the foundation for both ChatGPT and InstructGPT. ChatGPT is tailored to generate dialogic responses across diverse topics, while InstructGPT is designed to provide detailed answers to specific user prompts. Both models, although sharing a foundational architecture, have been fine-tuned using different methodologies and datasets to cater to their respective purposes. GPT-4, though specifics are not fully disclosed, is recognized to have a larger scale than its predecessor GPT-3.5, indicating improvements in model parameters and training data scope (14–16).

When evaluated on USMLE questions, GPT-4 outperformed all other models with accuracy rates of 80-86%, 81-89% and 81-90% in Step1, Step2 and Step3, respectively. ChatGPT also had relatively good results, outperforming GPT-3, instructGPT and GPT-3.5 with accuracy rates of 41-75%, 49-61% and 55-68% in Step1, Step2 and Step3, respectively (**Tables 3-5**).

## Questions with Media Elements

Some of the USMLE questions use media elements such as graphs, images, and charts (14.4% and 13% of questions in the Self-assessment and Sample exam, accordingly) (17). While Glison et al. and Kung et al. excluded these questions because these elements do not get passed to the model, Nori et al. included them in the evaluation. They found that while GPT-4 performs best on text-only questions, it still performs well on questions with media elements, with 68-79% accuracy, despite not being able to see the relevant images. GPT-3.5 was found to have a 41-53% accuracy in the media elements containing questions.

## Prompting methods

Different prompting methods were used to test the performance of the LLMs. Examples of prompts are shown in **Table 6**.

Kung et al. (18) tested three prompting options. In the Open-Ended format, answer choices were eliminated, and variable lead-in interrogative phrases were incorporated to mirror natural user queries. In the Multiple-Choice Single Answer without Forced Justification format, USMLE questions were reproduced exactly. The third format, Multiple Choice Single Answer with Forced Justification, required ChatGPT to provide rationales for each answer choice.

Glison et al. (16) used only multiple choice prompting.

Nori et al. (17) also used multiple choice questions but tested both zero-shot and few-shot prompting. Zero-shot prompting requires a model to complete tasks without any prior examples, utilizing only the knowledge gained from pre-training. On the other hand, few-shot prompting provides the model with limited examples of the task at hand before execution. The model is expected to generalize from these examples and accurately perform the task.

Overall, the performance of the models was slightly affected by the prompting method, with open-ended prompting showing better results than multiple choice, and 5-shot prompting giving better results than zero-shot.

**Performance assessment**

All papers assessed accuracy of the models in answering USMLE questions. Two studies also included qualitative assessment of the answers and explanations provided by the LLMs. Glison et al. (16) assessed each answer for logical reasoning (identification of the logic in the answer selection), use of information internal to the question, and use of information external to the question. ChatGPT was reported to use information internal to the question in 97% of questions. The use of information external to the question was higher in correct (90-93%) answers than incorrect answers (48-63%). In addition, every incorrect answer was labeled for the reason of the error: logical error (the response uses the pertinent information but does not translate it to the correct answer), information error (did not identify the key information needed) or statistical error

(an arithmetic mistake). Logical errors were the most common, found in 42% of incorrect answers (16). Kung et al. (18) evaluated each output for concordance and insight, by two physician reviewers. A high concordance of 94.6% was found across ChatGPT's answers. ChatGPT produced at least one significant insight in 88.9% of questions. The density of insight contained within the explanations provided by ChatGPT was significantly higher in questions answered accurately than in incorrect answers.

## Discussion

This review provides a comparative analysis of LLMs performance on USMLE questions. While GPT-4 secured accuracy rates within the 80-90% range, ChatGPT demonstrated competent results, outpacing the capabilities of previous models, GPT-3, InstructGPT, and GPT-3.5.

The results of this review underscore that the main factor that affects performance, are the inherent capabilities of the LLM. Other factors, including various prompting methods, incorporation of questions that include media, and variability in question sets had secondary roles. This observation emphasizes the priority of advancing core AI model development to ensure better accuracy and utility in complex sectors like healthcare. As we consider the integration of LLMs into medical domains, it's crucial to recognize that while optimizing external parameters can contribute to performance tweaks, the most profound improvements lie in the refinement of the model's core capabilities.

The proficiency of these LLMs on a rigorous and foundational examination such as the USMLE provides a compelling indication of these models' potential role in the medical domain. GPT-4's ability to achieve high accuracy levels signifies the progression and maturation of AI's capabilities in deciphering complex medical knowledge. Such advancements could be pivotal in assisting healthcare professionals, improving diagnostic accuracy, and facilitating medical education. However, while the results are promising, it is crucial to approach the integration of LLMs in medical practices with caution. The USMLE's textual nature might not encompass the

entire scope of clinical expertise, where skills like patient interaction (19), hands-on procedures, and ethical considerations play an important role.

Prompting is considered to hold a significant role in shaping the performance of LLMs when answering queries (20,21). This review demonstrates that the way questions are structured can subtly influence the responses generated by these models. Notably, open-ended prompting has a slight edge over the standard multiple-choice format. This suggests that LLMs might have a nuanced preference when processing information based on the context they're provided. Moreover, the marginally better outcomes with 5-shot prompting compared to zero-shot hint at the LLMs' capacity to adjust and produce informed answers when given a few guiding examples. Though these differences are subtle and may not dramatically change the overall performance, they provide valuable insights into the optimization of interactions with LLMs.

Two sets of formal USMLE questions were utilized in the studies reviewed. Both GPT-3.5 and GPT-4 exhibited superior performance on the Sample exam compared to the self-assessment. While the Sample exam is publicly accessible, the self-assessment can only be obtained through purchase. This raises the possibility that the higher accuracy is derived from previous encounters of the models with questions from the Sample exam. Nori et al. (15). developed an algorithm to detect potential signs of data leakage or memorization effects. This algorithm is designed to ascertain if specific data was likely incorporated into a model's training set. Notably, this method did not detect any evidence of training data memorization in the official USMLE datasets, which include both the self-assessment and the Sample exam. However, it is important to note that while the algorithm demonstrates high precision, its recall remains undetermined. Therefore, the extent to which these models might have been exposed to the questions during their training remains inconclusive.

In surveying the expansive literature on the application of LLMs in healthcare, it is noteworthy that only three studies are directly comparable in their use of formal question sets. This observation highlights the potential disparity in evaluation methods and emphasizes the need for standardized benchmarks in assessing LLMs′ medical proficiency. The USMLE stands as a primary metric for evaluating medical students and residents in the U.S., and its role as a benchmark for LLMs warrants careful consideration. While it offers a structured and recognized platform, it is crucial to contemplate whether such standardized tests can fully encapsulate the depth and breadth of LLMs′ capabilities in medical knowledge. Looking ahead there is a pressing need for research that delves into alternative testing mechanisms, ensuring a comprehensive and multidimensional evaluation of LLMs in the realm of healthcare.

This systematic review has several limitations. First, the studies reviewed primarily focused on multiple choice questions, which, although a prevalent format in the USMLE and an accepted method for assessing medical knowledge among students and clinicians, may not fully capture the complexity of real-world medical scenarios. Actual clinical cases often present with complexities and subtleties that might not strictly align with textbook descriptions. Hence, when contemplating the applicability of LLMs in a clinical setting, it is vital to recognize and account for this disparity. Secondly, our review intentionally excluded studies that examined other datasets used for USMLE preparation, such as MedQA. This decision was made to maintain consistency in the comparison of question sets. However, there is a wide array of research that evaluates various other LLMs using diverse question sets that mimic USMLE questions and measure medical proficiency. The exclusion of these studies potentially limits the comprehensiveness of our insights into the capabilities of LLMs in the medical domain.

To conclude, this systematic review presents a detailed analysis of the performance capabilities of LLMs on the USMLE. The primary influence of the core model′s capabilities over external factors, as highlighted in our findings, underscores the importance of continual advancements in

AI model development targeted to the medical field. While the USMLE is a reputable metric for assessing medical knowledge, it might not wholly capture the diverse challenges of clinical practice. Future research should strive for a broader range of benchmarks, transcending traditional testing systems, to fully understand the potential and boundaries of LLMs in healthcare.

**Figure 1**

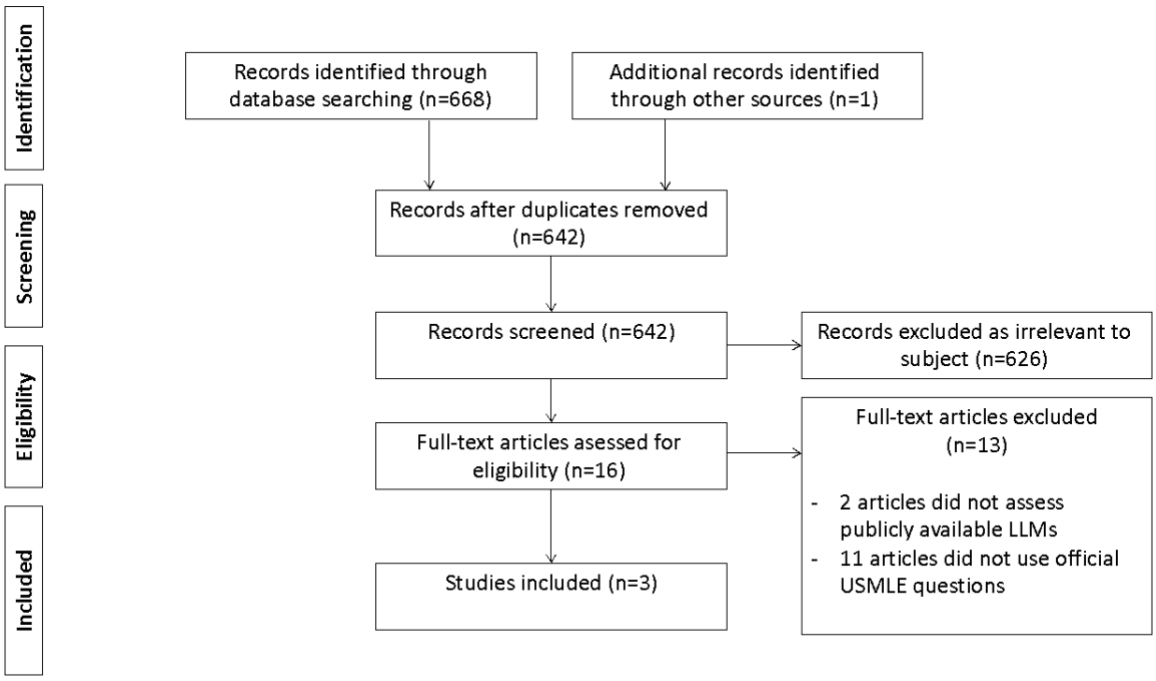


**Figure 1.** Flow diagram of the search and inclusion process in the study. The study followed the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) guidelines. USMLE = United States Medical Licensing Examination. LLMs = Large Language Models.

Table 1. Publications reporting on performance on LLMs in USMLE questions.

| Authors | Title | Journal | Year | Month |
|---|---|---|---|---|
| Kung et al. | Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models | PLOS digital health | 2023 | February |
| Glison et al. | How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment | JMIR medical education | 2023 | August |
| Nori et al. | Capabilities of GPT-4 on Medical Challenge Problems | arXiv | 2023 | April |

Table 2. Parameters evaluated in the publications reporting on performance on LLMs in USMLE questions.

| **Study (Ref)** | **Model(s) tested** | **USMLE step(s) evaluated** | **Questions sets used** | **Prompting Methodology** | **Images and graphs** | **Performance metric(s) used** |
|---|---|---|---|---|---|---|
| Kung et al. (18) | ChatGPT | Step 1,2,3 | Sample exam* | Open-ended, Multiple choice, Multiple choice with forced justification | excluded | Accuracy, concordance and insight scores by two physicians |
| Glison et al. (16) | ChatGPT, GPT-3, InstructGPT | Step 1,2 | Sample exam* | Multiple choice | excluded | Logical justification, presence of information internal to the question, presence of information external to the question |
| Nori et al. (17) | GPT-4, GPT-3.5 | Step 1,2,3 | Sample exam*, Self-assessment** | Open ended template– zero-shot, 5-shot | Included | - |

* Sourced from the official USMLE website
** Sourced from the official NMBE website

Table 3. LLM's performance (%) on USMLE Step1

| Ref. | Data Set | Prompting | GPT-3 | InstructGPT | GPT-3.5 | ChatGPT | GPT-4 |
|---|---|---|---|---|---|---|---|
| (18) | Sample exam | Open-ended<br>Multiple choice<br>Multiple choice with forced justification | | | | 75/45.4*<br>55.8/36.1*<br>64.5/41.2* | |
| (16) | Sample exam | Multiple choice | 25.3 | 51.7 | | 64.4 | |
| (17) | Self-assessment | Zero-shot<br>5-shot | | | 49.62<br>54.22 | | 83.46<br>85.21 |
| | Sample exam | Zero-shot<br>5-shot | | | 51.26<br>52.10 | | 80.67<br>85.71 |

* With indeterminate responses included

Table 4. LLM's performance on USMLE Step2

| Ref. | Data Set | Prompting | GPT-3 | InstructGPT | GPT-3.5 | ChatGPT | GPT-4 |
|---|---|---|---|---|---|---|---|
| (18) | Sample exam | Open-ended<br>Multiple choice<br>Multiple choice with forced justification | | | | 61.5/54.1*<br>59.1/56.9*<br>52.4/49.5* | |
| (16) | Sample exam | Multiple choice | 18.6 | 52.9 | | 57.8 | |
| (17) | Self-assessment | Zero-shot<br>5-shot | | | 48.12<br>52.75 | | 84.75<br>89.50 |
| | Sample exam | Zero-shot<br>5-shot | | | 60.83<br>58.33 | | 81.67<br>83.33 |

* With indeterminate responses included

Table 5. LLM's performance on USMLE Step3

| Ref. | Data Set | Prompting | GPT-3.5 | ChatGPT | GPT-4 |
|---|---|---|---|---|---|
| (18) | Sample exam | Open-ended<br>Multiple choice<br>Multiple choice with forced justification | | 68.8/61.5*<br>61.3/55.7*<br>65.2/59.8* | |
| (17) | Self-assessment | Zero-shot<br>5-shot | 50.00<br>53.41 | | 81.25<br>83.52 |
| | Sample exam | Zero-shot<br>5-shot | 58.39<br>64.96 | | 89.78<br>90.71 |

* With indeterminate responses included

Table 6. USMLE questions prompting.

| Prompting | Example | Ref. |
|---|---|---|
| Open ended | <question without multiple choice answers><br>"What would be the patient's diagnosis based on the information provided?" | (18) |
| Multiple choice | <question with multiple choice answers> | (16,18) |
| Multiple choice with forced justification | <question with multiple choice answers><br>"Explain your rationale for each choice"/ "Why are the other choices incorrect?" | (18) |
| Zero-shot | "The following are multiple choice questions (with answers) about medical knowledge."<br>"**Question: **" <question with multiple choice answers><br>"**Answer: **" | (17) |
| 5-shot | "The following are multiple choice questions (with answers) about medical knowledge."<br><few shot examples><br>"**Question: **" <question with multiple choice answers><br>"**Answer: **" | (17) |

Bibliography


1. About the USMLE | USMLE [Internet]. [cited 2023 Aug 2]. Available from: https://www.usmle.org/about-usmle

2. Lombardi CV, Chidiac NT, Record BC, Laukka JJ. USMLE step 1 and step 2 CK as indicators of resident performance. BMC Med Educ. 2023 Jul 31; 23(1): 543.

3. Ozair A, Bhat V, Detchou DKE. The US Residency Selection Process After the United States Medical Licensing Examination Step 1 Pass/Fail Change: Overview for Applicants and Educators. JMIR Med Educ. 2023 Jan 6; 9: e37069.

4. Chaudhry HJ, Katsufrakis PJ, Tallia AF. The USMLE Step 1 Decision: An Opportunity for Medical Education and Training. JAMA. 2020 May 26; 323(20): 2017–8.

5. Grunhut J, Marques O, Wyatt ATM. Needs, Challenges, and Applications of Artificial Intelligence in Medical Education Curriculum. JMIR Med Educ. 2022 Jun 7; 8(2): e35587.

6. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern Med. 2023 Jun 1; 183(6): 596.

7. Sahni NR, Carrus B. Artificial Intelligence in U.S. Health Care Delivery. Drazen JM, Kohane IS, Leong TY, editors. N Engl J Med. 2023 Jul 27; 389(4): 348–58.

8. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023 Jul 13; 619(7969): 357–62.

9. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. Npj Digit Med. 2023 Jul 6; 6(1): 1–6.

10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29; 372: n71.

11. Prepare for Your Exam | USMLE [Internet]. [cited 2023 Aug 7]. Available from: https://www.usmle.org/prepare-your-exam

12. Taking a Self-Assessment | NBME [Internet]. [cited 2023 Aug 7]. Available from: https://www.nbme.org/examinees/self-assessments

13. OpenAI Platform [Internet]. [cited 2023 Aug 2]. Available from: https://platform.openai.com

14. GPT-4 [Internet]. [cited 2023 Aug 7]. Available from: https://openai.com/research/gpt-4

15. Introducing ChatGPT [Internet]. [cited 2023 Aug 7]. Available from: https://openai.com/blog/chatgpt

16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8; 9: e45312.

17. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. ArXiv Prepr ArXiv230313375. 2023;

18. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb; 2(2): e0000198.

19. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, et al. Large Language Models (LLMs) and Empathy – A Systematic Review [Internet]. medRxiv; 2023 [cited 2023 Aug 10]. p. 2023.08.07.23293769. Available from: https://www.medrxiv.org/content/10.1101/2023.08.07.23293769v1

20. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023; 1–9.

21. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Internet]. arXiv; 2023 [cited 2023 Aug 7]. Available from: http://arxiv.org/abs/2201.11903

22. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? 2022 [cited 2023 Jul 20]; Available from: https://arxiv.org/abs/2207.08143