Cureus Journal of Computer Science

Open Access Original Article

Analyzing the Accuracy of Large Language Models in United States Medical Licensing Exam Social-Science Preparation Question Banks

Received 07/09/2025
Review began 07/30/2025
Review ended 08/22/2025
Published 09/01/2025

DOI:
https://doi.org/10.7759/s44389-025-09114-9

Raika Bourmand [1] , Sofia E. Olsson [2] , Fiza Z. Baloch [1] , Akhila Sonti [1] , Arman Fijany [3]

1. Medicine, Burnett School of Medicine at TCU, Fort Worth, USA 2. Otolaryngology, UC Davis Medical Center, Sacramento, USA 3. Plastic Surgery, UC Davis Medical Center, Sacramento, USA

Corresponding author: Raika Bourmand, raika.bourmand@tcu.edu

Abstract

Introduction: Artificial intelligence (AI) emergence has changed the medical education landscape. The United States Medical Licensing Exams increasingly include questions in the “Social Sciences (Ethics/Legal/Professional)” domain, which often require reasoning through complex scenarios. This study evaluates three AI platforms in answering such preparation questions.

Methods: Social-science questions from UWorld and Amboss were accumulated for Steps 1 and 2. Multiple-choice questions were entered into ChatGPT, Gemini, and Perplexity, yielding a correct/incorrect response. Percentages of correct responses were compared between platforms and to student averages. Analysis of variance and t-tests conducted determined statistical significance, the upper threshold being $P = 0.05$. As this study focused on quantitative performance, outcomes were limited to accuracy on multiple-choice items, rather than direct measures of empathetic reasoning.

Results: One hundred nine UWorld and 63 Amboss questions were available for Step 1. A total of 189 UWorld and 185 Amboss questions were available for Step 2. Google Gemini had the highest accuracy for Step 1 (86.6%), while ChatGPT had the highest accuracy for Step 2 (83.6%). All platforms outperformed the student average for Step 1 (68.5%) and Step 2 (68.9%), with ChatGPT and Gemini doing so significantly for Step 1 ($p < 0.01$), and ChatGPT doing so significantly for Step 2 ($p < 0.01$).

Conclusion: It is critical to understand whether empathetic thinking can be replicated by technology or prepare students. This study highlights the ability for AI to solve ethical dilemmas that students may struggle with.

Categories: AI applications, AI Ethics and Fairness, AI/ML-based decision support systems
Keywords: artificial intelligence, medical ethics, medical education, large language model(llm), exam preparation

Introduction

The rapid advancement and widespread adoption of artificial intelligence (AI) and large language models (LLMs) is transforming multiple industries including various aspects of medicine. The AI market, valued at $50 billion in 2023, is projected to grow remarkably, reaching $826 billion by 2030 [1]. With the democratization of AI tools, these emerging technologies may offer new ways to support students navigating medical training, including preparation for pivotal exams such as the United States Medical Licensing Examination (USMLE). As medicine adapts to this technological shift, the role of LLMs in shaping how future physicians learn, reason, and make decisions is becoming a central question.

Despite the promise of AI in medical education, its integration has been met with mixed responses [2-4]. A 2023 study by Alkhaaldi et al. [2] found that only 20.4% of surveyed residency applicants endorsed using AI to complete written assessments, and 9.4% used AI for clinical work. However, the majority of applicants expressed interest in utilizing AI during residency to explore new medical topics and exam preparation [2]. It appears that AI use is increasing drastically in medical education, with a 2024 study by Zhang et al. [3] finding that nearly half of United States medical students were using AI tools like ChatGPT to assist in their studies, particularly for writing, revising, and summarizing content. Similarly, a 2022 survey by Pucchio et al. [4] revealed that 94% of medical students believe AI will become more common in the future, with 84% agreeing that it will improve medicine. A majority also advocated for formal AI education in medical curricula to optimize the use of emerging technologies [4].

In the realm of medical education, LLMs like Generative Pre-trained Transformer 4 (GPT-4) have demonstrated the ability to effectively answer USMLE board-style questions both for the Step 1 and Step 2 examinations [5]. With an increased number of questions on ethical scenarios in the USMLE [6], there is a question of whether LLMs have the ability to empathetically and ethically reason through complex social scenarios at the level comparable to medical students. These questions are important as they are designed

How to cite this article
Bourmand R, Olsson S E, Baloch F Z, et al. (September 01, 2025) Analyzing the Accuracy of Large Language Models in United States Medical Licensing Exam Social-Science Preparation Question Banks . Cureus J Comput Sci 2 : es44389-025-09114-9. DOI https://doi.org/10.7759/s44389-025-09114-9

to assess a physician's ability to navigate complex moral dilemmas with empathy and strong ethical judgment. These are skills included in a majority of medical curricula, which have been identified as essential for providing compassionate patient care [7]. UWorld is one of the most used online preparation resources for standardized medical examinations, with over 90% of medical students choosing this resource for exam preparation [8]. UWorld has been previously shown by multiple studies to be a predictive factor for USMLE exam performance [9-10]. Similarly, Amboss has previously been associated with positive learning outcomes for medical students [11], with every two out of three United States medical students using the online platform to prepare for their examinations [12]. In preparation for USMLE exams, previous literature has found that ChatGPT is capable of performing at or near the passing threshold for Step 1, Step 2 Clinical Knowledge (CK), and Step 3 questions, although this study did not specify how well specific test sections were performed, such as the ethics discipline [5].

Of the LLMs in the market, ChatGPT hosts nearly 300 million weekly active users worldwide, 70 million of which are from the United States. ChatGPT also accounts for nearly 63% of the market share of all AI tools [13]. A recent study found that 96% of medical students had heard of ChatGPT with 52% using it for medical school coursework [14]. Likewise, Google Gemini has an average of nearly 275 million monthly visits, with the highest share of traffic being from the United States [15]. Perplexity has 10 million active monthly users, with 45 million visits in December 2023. The United States is also in the top three user bases of Perplexity [16]. A major commonality and important consideration of these three LLMs is their affordability as free-to-use platforms, thereby making them more accessible for medical students [4,17-18]. When comparing the three models in terms of content generation, a recent study found that ChatGPT is best at processing textual data, in comparison to Gemini and Perplexity who are capable of processing both text and visual-based data [19].

A previous study on 80 board-style questions found AI to be effective in answering these ethically charged questions, with GPT-4 outperforming ChatGPT [20]. The present study aims to build on this body of research by analyzing the ethical decision-making capacity of a variety of generative AI tools with the largest question sample size to date. By comparing the responses of multiple LLMs, this study seeks to provide a comprehensive understanding of AI platform strengths and limitations in addressing the ethical challenges faced by medical professionals. In doing so, this work contributes not only to the evaluation of AI as a practical study tool, but also to the broader conversation about how LLMs may reshape the ethical reasoning, empathy, and clinical judgment that define the practice of medicine in the 21st century.

Materials And Methods

Input source

Three LLMs were used in this study: ChatGPT-4o, Google Gemini 2.0 Flash, and Perplexity free [21-23]. ChatGPT-4o is a transformer-based autoregressive language model trained on large-scale textual data to predict and generate human-like text [21]. Google Gemini 2.0 Flash is a multimodal LLM that leverages both text and image inputs with transformer-based attention mechanisms, optimized for general reasoning tasks [22]. Perplexity free is a retrieval-augmented LLM that combines pre-trained language generation with external knowledge sources to provide informative and contextually relevant responses [23]. All three models generate responses using probabilistic pattern recognition over learned language representations rather than explicit rule-based reasoning. They were chosen based on their large United States user base and free-to-use option.

All questions within the system of “Social Sciences (Ethics/Legal/Professional)” of UWorld and “Social Sciences” of Amboss were accumulated from the Step 1 and 2CK question banks [24]. Question collection was conducted between January 3-9, 2025, acknowledging that UWorld and Amboss periodically update their banks and that future replications may differ slightly depending on the versions available at the time of access. This totaled to four preparation question banks: 109 from UWorld Step 1, 189 for Uworld Step 2, 63 for Amboss Step 1, and 185 for Amboss Step 2 CK. Each question along with its multiple-choice answers was entered in a randomized order into each respective LLM model. The scoring consisted of allocating one point based on whether each question entered into the LLM model matched the correct answer indicated by UWorld or Amboss preparation banks. If the LLM model incorrectly identified the letter choice based on the options provided by the preparation bank, then it received a score of 0. Each question was only inputted once into the LLM models. This “one-shot” approach was chosen to provide a consistent and reproducible assessment of initial LLM accuracy, reflecting a standardized testing scenario. We acknowledge that it may disadvantage models capable of refining answers after follow-up prompts, which some students might use in practice. The number of points based on each LLMs correctness of questions within their respective question banks was summed, and divided by the total number of questions within that preparation bank (109 from UWorld Step 1, 189 for UWorld Step 2, 63 for Amboss Step 1, and 185 for Amboss Step 2 CK). This resulted in each of the four question banks having three percentage scores for ChatGPT-4o, Google Gemini 2.0 Flash, and Perplexity free.

A weighted average score for each LLM was also created based on their overall performance for either Step 1 preparation questions, across both UWorld and Amboss using the following formula: ((109*UWorld%

correct)+(63*Amboss% correct))/(109+63). Similarly, this was also done for Step 2, using the following formula: ((189*UWorld% correct)+(185*Amboss% correct))/(189+185). This resulted in each LLM having two weighted average scores, one for Step 1 and another for Step 2 preparation questions. Weighted averages were calculated for each LLM across UWorld and Amboss question banks to ensure proportional representation of question bank sizes, so that larger banks contribute appropriately more to the overall performance metric than smaller banks.

Average medical student percentage correct scores were gathered for each of the Social Science preparation question banks, based on the information provided from both UWorld and Amboss preparation resources. These averages reflect performance on practice questions and are not standardized across cohorts, nor are they equivalent to final USMLE exam performance. Weighted averages were calculated using the same formulas to provide a reference point for LLM performance.

To address the scope of question content, we note that the UWorld “Social Sciences (Ethics/Legal/Professional)” and Amboss “Social Sciences” categories are exam preparation resources modeled after, but not identical to, the NBME Social Sciences content outline. While they are widely used by medical students to prepare for USMLE examinations [8,12], they do not represent the official NBME item pool. Within our dataset, questions represented a mixture of ethical/empathetic, ethical/legal, and professional responsibility items. While a detailed subcategorical breakdown (e.g., proportion of empathetic vs. legal reasoning questions) was not performed in this study, we recognize this as a limitation and an opportunity for future work to more precisely categorize and analyze question types.

Statistical analysis

The percentage of correct responses per platform was calculated separately for Step 1 and Step 2, and overall accuracy scores were computed to analyze differences between AI platforms and question bank origins. UWorld and Amboss provide the average score among medical students who completed the same questions, and these scores were recorded for comparison with LLM accuracy.

A one-way analysis of variance (ANOVA) was performed using Microsoft Excel [24] to assess differences between LLMs for each question set, with p-values reported precisely (e.g., P = 0.60). Post-hoc pairwise comparisons were conducted using t-tests with Bonferroni correction to account for multiple comparisons and reduce Type I error. Effect sizes were calculated for both ANOVA (η^2) and t-tests (Cohen’s d) to quantify the magnitude of differences between LLMs and between LLMs and medical student averages. A significance threshold of P = 0.05 was applied.

Results And Discussion

Experimental setup

The questions were obtained using a new anonymized account for UWorld and Amboss preparation resources. A data table was created in Microsoft Excel [25] for each of the four question banks (UWorld Step 1, UWorld Step 2CK, Amboss Step 1, and Amboss Step 2 CK). This sheet was utilized for the purpose of scoring a question within each bank, either 1 point or 0 points across each LLM. A new private window was opened and utilized for each set of questions for the four preparation banks, and each LLM assessed. Each question was then copied and pasted one at a time into its respective LLM.

Step 1

This study yielded a total of 109 UWorld questions and 63 Amboss questions under the Step 1 social science/ethics discipline. As demonstrated in Table *1*, ChatGPT answered 82.6% (90/109) of UWorld questions correctly, while Google Gemini performed slightly better at 85.3% (93/109), and Perplexity performed with a 73.3% (80/109) correctness. ChatGPT answered 90.4% (57/63) of Amboss questions correctly, with Google Gemini similarly answering 88.8% (56/63) correctly. Perplexity had the lowest correct response rate at 73.0% (46/63).

Large language model	UWorld question bank percent correct	Amboss question bank percent correct	Weighted average of question banks percent correct	P-value
ChatGPT 4o	82.6	90.4	85.5	* <0.005
Google Gemini	85.3	88.8	86.6	* <0.005
Perplexity	73.3	73	73.2	p = 0.098
Average of all AI models	80.4	84.1	81.8	* <0.005
Average medical student score	66.7	71.6	68.5	

TABLE 1: USMLE Step 1 social science/ethics question banks correctness.

P-value indicating significant differences between average large language model score and average medical student score using t-tests (*P ≤ 0.05).

USMLE, United States Medical Licensing Examination

Step 2

This study yielded a total of 189 UWorld questions and 185 Amboss questions under the Step 2 social science/ethics discipline. ChatGPT had the highest correct response rate for Amboss preparation questions at 81.0% (150/185) (Table 2). Meanwhile, Google Gemini's correct response rate was 72.9% (135/185) and Perplexity was similar in performance with an accuracy rate of 70.2% (130/185). On average, ChatGPT answered 86.2% (163/189) of UWorld preparation questions correctly, while Google Gemini performed slightly higher at 87.3% (165/189). Perplexity performed the highest of all platforms at 87.8% (166/189) (Table 2).

Large language model	UWorld question bank percent correct	Amboss question bank percent correct	Weighted average of question banks percent correct	P-value
ChatGPT	86.2	81	83.6	*p = 0.047
Google Gemini	87.3	72.9	80.2	p = 0.199
Perplexity	87.8	70.2	79.1	p = 0.278
Average of all AI models	87.1	74.7	81.0	p = 0.154
Average medical student score	65.8	72.0	68.9	

TABLE 2: USMLE Step 2 social science/ethics question banks correctness

P-value indicating significant differences between average LLM score and average medical student score using t-tests (*P ≤ 0.05).

USMLE, United States Medical Licensing Examination

It was determined that Perplexity had the worst performance across both UWorld and Amboss question banks for both Step 1 and Step 2 preparation. ChatGPT was determined to be the most accurate LLM in answering Amboss questions for either Step 1 or 2 preparation. Google Gemini was the most accurate tool to use for UWorld Step 1 preparation.

ANOVA demonstrated no statistically significant differences between the UWorld and Amboss question banks for both Step 1 ($P = 0.60$, $\eta^2 = 0.004$) and Step 2 ($P = 0.35$, $\eta^2 = 0.004$). ANOVA determined that accuracy was significantly different between the three LLMs in the Step 1 combined question banks ($P < 0.001$, $\eta^2 = 0.12$). However, ANOVA did not identify a statistically significant difference between accuracy of the LLMs in the Step 2 question bank ($P = 0.08$, $\eta^2 = 0.007$).

Post-hoc t-tests with Bonferroni correction revealed that ChatGPT ($P = 0.01$, Cohen's $d = 0.61$) and Google Gemini ($P = 0.01$, Cohen's $d = 0.58$) were significantly more accurate than medical students in answering Step 1-level social science/ethics questions. The question-solving accuracy of Perplexity was not significantly different from medical students ($P = 0.10$, Cohen's $d = 0.18$). For Step 2, statistically significant differences were observed only between ChatGPT and medical students ($P = 0.01$, Cohen's $d = 0.48$), whereas differences for Gemini ($P = 0.199$, Cohen's $d = 0.12$) and Perplexity ($P = 0.278$, Cohen's $d = 0.10$) were not significant.

These results suggest that while performance among LLMs is largely comparable, ChatGPT consistently outperformed medical students with moderate effect sizes, whereas differences for other LLMs were smaller or non-significant.

The purpose of this study was to assess and compare the accuracy of three LLMs including ChatGPT, Google Gemini, and Perplexity, in solving board-style questions from the two most commonly used social science/ethics question banks used by medical students preparing for the USMLE. The results demonstrated notable differences in the accuracy of responses between the LLMs and average medical student scores, highlighting their ability to navigate complex empathy-based scenarios and potential for serving as tools for medical board exam preparation.

The consistent low performance of Perplexity across both question banks for Step 1 and Step 2 preparation suggests that it may not be the ideal LLM in preparing for complex social and ethical board-style questions. These findings are further supported by previous studies in which Perplexity was compared with three other LLMs in medical decision-making and was found to have the lowest mean accuracy [26]. Our results suggest that ChatGPT and Google Gemini may be better suited options for learners seeking guidance with ethical dilemmas in medicine. However, it is important to note that this study solely observed the accuracy of AI-generated responses to multiple choice questions, which may not fully encompass the multi-faceted approach to assessing the quality of an educational tool. When considering the possibility of using AI as a learning tool, additional factors must be considered, including quality of explanations generated by the LLM.

The ability of LLMs to respond to empathy-based prompts is of increasing importance. Medical students face an overwhelming amount of information to master, and LLM tools serve as valuable assistance in studying for board examinations. By assessing the reliability of LLMs in specific question banks, students can make informed decisions about which LLMs to incorporate into their study routines. LLMs are also becoming more prevalent in medical decision-making. A recent study comparing responses of ChatGPT and physicians to patient inquiries through social media found that ChatGPT's responses were viewed as more empathetic, emphasizing its potential to display human-like characteristics [27]. Therefore, it is essential that LLM can provide reliable responses and be used as an effective supplement.

The present study is the first of its kind to compare the performance of different LLMs in responding to complex ethical scenarios in the form of board-style questions. It is also the first to do so with two input sources including AMBOSS and UWorld, yielding the largest sample size of questions to date. A previous study assessed the ability of ChatGPT and GPT-4 in answering ethics-style questions, with an overall accuracy score of 62.5%; however, multiple AI platforms and a broad question sample were not assessed [28]. However, the notable difference in accuracy between the aforementioned study and our study poses the question of whether AI platforms have continued to strengthen neural networks relating to ethical questions in medicine over time. ChatGPT-4.5 [29] represents an important evolution in LLM capabilities, demonstrating improvements in intuition and scaling reasoning. Given its enhanced ability to process complex ethical dilemmas and provide empathetic responses, it may serve as a study aid in reflective learning regarding medical ethics and patient interactions. Future research should examine how ChatGPT-4.5 performs in open-ended or interactive cases as a comprehensive medical learning tool.

A limitation of this study is that, due to limited sample size, it was unable to further assess subgroups of these ethics questions such as difficulty ratings or related specialty. This may have provided insight into specific strengths or weaknesses in each LLM. The study also did not assess the quality of AI-generated explanations for their answers, which would provide more information on how each platform may be used as a learning tool for students. Additionally, students are learning when performing board-style question banks. Due to the role of UWorld and Amboss as study tools, they may present deflated medical student averages in comparison to student performance on the official USMLE. It is also important to note that each LLM was given only one attempt per question ("one-shot" prompting) in this study. While this approach reflects a standardized testing scenario and allows reproducible assessment of initial accuracy, it

may disadvantage models capable of refining answers through follow-up prompts. In practice, students may interact iteratively with LLMs, requesting explanations or clarifications, which could impact performance. This represents an opportunity for future research to evaluate multi-step or interactive prompting strategies.

Conclusions

This study provides valuable insights into the accuracy and effectiveness of different LLMs for medical exam preparation within the social science/ethics discipline. The dataset included multiple-choice questions from UWorld and Amboss question banks in the categories of ethical/legal and professional responsibility scenarios. The results suggest that ChatGPT and Gemini may both be strong supplemental resources for Step 1 question banks; however, ChatGPT may be the only strong supplemental LLM for Step 2 question banks. ChatGPT was shown to be the most reliable LLM for this style of question, while Perplexity was the least reliable; however, all the AI platforms studied outperformed medical students on these practice questions. Given the growing role of LLMs in medical education and decision-making, their ability to provide accurate responses in complex ethical scenarios is increasingly important. Future studies may explore LLM performance across diverse medical disciplines, assess the impact of continuous model updates, and examine the quality of explanations provided. Understanding these factors will help medical students make informed decisions about integrating AI tools into their study routines and ensure that these models serve as effective educational supplements.

Appendices

Link to Dataset: https://docs.google.com/spreadsheets/d/1rcF_-G6yWfGScrRdHPy-j8rg1tI6M3OwYLQ4YniCcms/edit?usp=sharing

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Raika Bourmand

Acquisition, analysis, or interpretation of data: Raika Bourmand, Sofia E. Olsson, Fiza Z. Baloch, Akhila Sonti, Arman Fijany

Drafting of the manuscript: Raika Bourmand, Sofia E. Olsson, Fiza Z. Baloch

Critical review of the manuscript for important intellectual content: Sofia E. Olsson, Akhila Sonti, Arman Fijany

Supervision: Arman Fijany

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

This article was previously presented as a poster at the University of Texas Medical Branch Research Forum on April 6, 2025 as well as posted to the Authorea preprint server on April 15, 2025, and is not pending full publication elsewhere.

References

1. Global AI Market Size 2030. (2024). Accessed: March 19, 2025: https://www.statista.com/forecasts/1474143/global-ai-market-size.
2. Alkhaaldi SMI, Kassab CH, Dimassi Z, Oyoun Alsoud L, Al Fahim M, Al Hageh C, Ibrahim H: Medical student experiences and perceptions of ChatGPT and artificial intelligence: Cross-sectional study. JMIR Medical Education. 2023, 9:e51302. 10.2196/51302
3. Zhang JS, Yoon C, Williams DKA, Pinkas A: Exploring the usage of ChatGPT among medical students in the

United States. Journal of Medical Education and Curricular Development. 2024, 11:10.1177/23821205241264695
4. Pucchio A, Rathagirishnan R, Caton N, et al.: Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Medical Education. 2022, 22:815. 10.1186/s12909-022-03896-5
5. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023, 2:e0000198. 10.1371/journal.pdig.0000198
6. ECFMG News | USMLE. (2019). Accessed: March 19, 2025: https://www.ecfmg.org/news/2019/11/06/usmle-step-1-and-step-2-ck-content-changes-scheduled-for-mid-2020/.
7. Ethics in Health Care: Improving Patient Outcomes. Tulane University. (2023). Accessed: March 19, 2025: https://publichealth.tulane.edu/blog/ethics-in-healthcare/.
8. UWorld Medical frequently asked questions. (2024). Accessed: February 9, 2025: https://www.ama-assn.org/medical-students/usmle-step-1-2/uworld-medical-frequently-asked-questions.
9. Parry S, Pachunka J, Beck Dallaghan GL: Factors predictive of performance on USMLE Step 1: Do commercial study aids improve scores?. Medical Science Educator. 2019, 29:667-672. 10.1007/s40670-019-00722-4
10. Giordano C, Hutchinson D, Peppler R: A predictive model for USMLE Step 1 scores. Cureus. 2016, 8:e769. 10.7759/cureus.769
11. Bientzle M, Hircin E, Kimmerle J, Knipfer C, Smeets R, Gaudin R, Holtz P: Association of online learning behavior and learning outcomes for medical students: Large-scale usage data analysis. JMIR Medical Education. 2019, 5:e13529. 10.2196/13529
12. Find out about your students' usage on AMBOSS. (2025). Accessed: February 9, 2025: https://www.amboss.com/us/institutions/usage-report.
13. Backlinko ChatGPT Statistics 2024: How Many People Use ChatGPT?. (2024). Accessed: March 19, 2025: https://backlinko.com/chatgpt-stats.
14. Ganjavi C, Eppler M, O'Brien D, et al.: ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLoS Digital Health. 2024, 3:e0000596. 10.1371/journal.pdig.0000596
15. Kumar N: How Many People Use Gemini in 2025 (Users & Demographics). (2025). https://www.demandsage.com/google-gemini-statistics/.
16. Perplexity AI: The Game-Changer in Conversational AI and Web Search. (2024). Accessed: March 19, 2025: https://originality.ai/blog/perplexity-ai-statistics.
17. ChatGPT. (2024). Accessed: March 19, 2025: https://openai.com/chatgpt/overview/.
18. Gemini API. (2024). Accessed: March 19, 2025: https://ai.google.dev/pricing#1_5flash.
19. Shukla M, Goyal I, Gupta B, Sharma J: A comparative study of ChatGPT, Gemini, and Perplexity. International Journal of Innovative Research in Computer Science & Technology. 2024, 12:10-15. 10.55524/ijircst.2024.12.4.2
20. Guerra GA, Hofmann H, Sobhani S, et al.: GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurgery. 2023, 179:e160-e165. 10.1016/j.wneu.2023.08.042
21. OpenAI ChatGPT. (2025). Accessed: February 9, 2025: https://chatgpt.com/.
22. Google. (2025). Accessed: February 9, 2025: https://gemini.google.com/app.
23. Perplexity. (2025). Accessed: February 9, 2025: https://www.perplexity.ai/.
24. UWorld | Test Prep for NCLEX, SAT, ACT, MCAT, USMLE. (2025). Accessed: February 9, 2025: https://www.uworld.com/?srsltid=AfmBOoq2h9O_xQqbuLIrX6EHqWkJLByFvfCAFWJRUEBSanbx-WVWGriK.
25. Microsoft Excel Online. (2025). Accessed: February 9, 2025: https://excel.cloud.microsoft/en-us/.
26. Uppalapati VK, Nag DS: A comparative analysis of AI models in complex medical decision-making scenarios: Evaluating ChatGPT, Claude AI, Bard, and Perplexity. Cureus. 2024, 16:e52485. 10.7759/cureus.52485
27. Ayers JW, Poliak A, Dredze M, et al.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine. 2023, 183:589-596. 10.1001/jamainternmed.2023.1838
28. Brin D, Sorin V, Vaid A, et al.: Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Scientific Reports. 2023, 13:16492. 10.1038/s41598-023-43436-9
29. OpenAI. (2025). Accessed: March 19, 2025: https://openai.com/index/introducing-gpt-4-5/.