

Evaluating AI Models for the National Pre-Medical Exam in India: A Head-to-Head Analysis of GPT-3.5, GPT-4, and Bard

Faiza Farhat, Beenish Moalla Chaudhry, Mohammad Nadeem, Shahab Saquib
Sohail, Dag Øivind Madsen

Submitted to: JMIR Medical Education
on: August 02, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Evaluating AI Models for the National Pre-Medical Exam in India: A Head-to-Head Analysis of GPT-3.5, GPT-4, and Bard

Faiza Farhat¹; Beenish Moalla Chaudhry²; Mohammad Nadeem³; Shahab Saquib Sohail⁴; Dag Øivind Madsen⁴

¹Department of Zoology Aligarh Muslim University Aligarh IN

²School of Computing and Informatics The University of Louisiana Lafayette US

³Department of Computer Science Aligarh Muslim University Aligarh IN

⁴USN School of Business University of South-Eastern Norway Hønefoss NO

Corresponding Author:

Dag Øivind Madsen

USN School of Business

University of South-Eastern Norway

Bredalsveien 14

Hønefoss

NO

Abstract

Background: Large language models (LLMs) have revolutionized Natural Language Processing (NLP) with their ability to generate human-like text through extensive training on large datasets. These models, including GPT-3.5, GPT-4, and Bard, find applications beyond NLP, attracting interest from academia and industry. Students are actively leveraging LLMs to enhance learning experiences and prepare for high-stakes exams, such as the National Eligibility cum Entrance Test (NEET) in India.

Objective: This comparative analysis aims to evaluate the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions.

Methods: In this paper, we evaluate the performance of the three mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET 2023 exam. The questions of NEET were provided to these AI models, and the responses were recorded and compared against the correct answers from the official answer key. Precision is used to evaluate the performance of all three models.

Results: It is evident that GPT-4 passed the entrance test with flying colors (43%), showcasing exceptional performance. On the other hand, GPT-3.5 managed to qualify but with a significantly lower score (21%). However, Bard (16%) failed to meet the qualifying criteria and did not pass the test. GPT-4 demonstrated consistent superiority over Bard and GPT-3.5 in all three subjects. Specifically, GPT-4 achieved accuracy rates of 72.5% in Physics, 44.44% in Chemistry, and 50.5% in Biology. Conversely, GPT-3.5 attained an accuracy rate of 45% in Physics, 33.33% in Chemistry, and 34.34% in Biology.

Conclusions: The study's findings provide valuable insights into the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions. GPT-4 emerged as the most accurate model, highlighting its potential for educational applications. The results underscore the suitability of LLMs for high-stakes exams and their positive impact on education. Additionally, the study establishes a benchmark for evaluating and enhancing LLMs' performance in educational tasks, promoting responsible and informed use of these models in diverse learning environments.

(JMIR Preprints 02/08/2023:51523)

DOI: <https://doi.org/10.2196/preprints.51523>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/51523>, the full text will be available to the public.



Original Manuscript

Evaluating AI Models for the National Pre-Medical Exam in India: A Head-to-Head Analysis of GPT-3.5, GPT-4, and Bard

Faiza Farhat¹, Beenish Moalla Chaudhry², Mohammad Nadeem³, Shahab Saquib Sohail⁴, Dag Øivind Madsen^{5,*}

¹Department of Zoology, Aligarh Muslim University, Aligarh, 202002, India

²School of Computing and Informatics, The University of Louisiana, Lafayette, LA 70504

³Department of Computer Science, Aligarh Muslim University, Aligarh 202002, India

⁴Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi 110062, India

⁵School of Business, University of South-Eastern Norway, 3511 Hønefoss, Norway

Abstract:

Background: Large language models (LLMs) have revolutionized Natural Language Processing (NLP) with their ability to generate human-like text through extensive training on large datasets. These models, including GPT-3.5, GPT-4, and Bard, find applications beyond NLP, attracting interest from academia and industry. Students are actively leveraging LLMs to enhance learning experiences and prepare for high-stakes exams, such as the National Eligibility cum Entrance Test (NEET) in India.

Objective: This comparative analysis aims to evaluate the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions.

Methods: In this paper, we evaluate the performance of the three mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET 2023 exam. The questions of NEET were provided to these AI models, and the responses were recorded and compared against the correct answers from the official answer key. Precision is used to evaluate the performance of all three models.

Results: It is evident that GPT-4 passed the entrance test with flying colors (43%), showcasing exceptional performance. On the other hand, GPT-3.5 managed to qualify but with a significantly lower score (21%). However, Bard (16%) failed to meet the qualifying criteria and did not pass the test. GPT-4 demonstrated consistent superiority over Bard and GPT-3.5 in all three subjects. Specifically, GPT-4 achieved accuracy rates of 72.5% in Physics, 44.44% in Chemistry, and 50.5% in Biology. Conversely, GPT-3.5 attained an accuracy rate of 45% in Physics, 33.33% in Chemistry, and 34.34% in Biology.

Conclusion: The study's findings provide valuable insights into the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions. GPT-4 emerged as the most accurate model, highlighting its potential for educational applications. The results underscore the suitability of LLMs for high-stakes exams and their positive impact on education. Additionally, the study establishes a benchmark for evaluating and enhancing LLMs' performance in educational tasks, promoting responsible and informed use of these models in diverse learning environments.

Introduction:

Large language models (LLMs) are potent Natural Language Processing (NLP) tools, excelling in a range of AI tasks, from news writing to product descriptions. They have garnered widespread attention across academia and industry [1, 2], going beyond the scope of natural language processing (NLP) into tasks related to healthcare [3], neuroscience [4], philosophy [5], marketing and finance [6, 7], sociology [8], education, and others [9, 10]. The development of LLMs and chatbots is experiencing an upsurge, with established companies and emerging startups actively engaged in their creation [11], catering to general or specific purposes [12]. Prominent examples include GPT-3.5

(developed by OpenAI), GPT-4 (also by OpenAI), and Bard (developed by Google) [13]. Other notable examples are BlenderBot, Galactica, LLaMA (FAIR) [14], Alpaca (Stanford), BloombergGPT [15], Chinchilla (DeepMind), and Palm [16], heralding the emergence of even more chatbots in the future [12].

The public release of ChatGPT in November 2022 and Bard in March 2023 has garnered significant attention due to their general purpose and flexible nature. ChatGPT[17], built on the GPT-3.5 architecture, has become popular for its remarkable ability to generate coherent and human-like responses. GPT-4.0 represents the latest iteration, incorporating enhanced language generation and improved multiturn conversation handling. Both GPT-3.5 and GPT-4.0 have been specifically trained to interact with users in a conversational manner, maintaining context, handling follow-up questions, and even correcting themselves. Bard, on the other hand, leverages Google's LaMDA [18], enabling it to handle a diverse range of language-related tasks and provide in-depth information.

In educational settings, students are utilizing LLMs such as Bard, GPT-3.5, and GPT-4 to enrich their daily learning experiences [20, 21]. They aid students in test preparation, offer research assistance, and contribute to their overall performance improvement and knowledge acquisition [22]. It has been observed that LLMs, despite their impressive performance, can sometimes generate text that includes fabricated or incorrect information [13, 23]. Consequently, researchers have directed their attention toward investigating the test-taking capabilities of different LLMs. Numerous research studies have delved into the assessment of GPT-3.5's efficacy in multiple-choice exams in higher education domains [24]. Some investigations have specifically focused on ChatGPT's test-taking performance in diverse professional fields, including business [25], accounting [26], law [27], and medicine [28]. In the medical realm, authors in [29] examined ChatGPT's competence in tackling the Medical College Admissions Test (MCAT), which serves as a prerequisite for admission to most medical schools in the United States. In [30] and [31], authors have scrutinized ChatGPT's aptitude in the United States Medical Licensing Examination (USMLE), while [32] conducted a comparative study of GPT-3.5 and GPT-4's performance in the Ophthalmic Knowledge Assessment Program (OKAP) Exam. Additionally, [33] undertook a comparison of GPT-3.5, GPT-4, and Google Bard, utilizing questions specifically prepared for neurosurgery oral board examinations. Similarly, [28] investigated ChatGPT's performance in several medical topics, namely, American Heart Association (AHA), Advanced Cardiovascular Life Support (ACLS) and Basic Life Support (BLS) exams.

Our primary objective is to bridge this knowledge gap by undertaking a comparative analysis of three notable chatbots: GPT-3.5 & GPT-4 by OpenAI, and Bard by Google, for a standardized medical school exam known as National Eligibility cum Entrance Test (NEET). NEET is a competitive entrance exam in India for MBBS and BDS programs in both government and private colleges. Introduced in 2013 by the Medical Council of India, NEET replaced various state-level and institution-specific tests to standardize medical admissions. Since 2019, National Testing Agency (NTA) has been responsible for conducting and supervising the NEET. The exam comprises a total of 200 multiple-choice questions aimed at testing knowledge, understanding, and aptitude in four subjects: Physics, Chemistry, Botany and Zoology. Candidates can only attempt a maximum of 45 questions per subject for a total of 180 out of 200 questions. Correct answers are awarded four (4) points, while each incorrect response leads to one (1) point deduction. Candidates are allotted three hours to complete the examination. To qualify for admission to a medical school, candidates must obtain a minimum or cutoff score, which can change year by year. The cutoff score for NEET 2023 was 137 out of 720. In 2023, over 2.03 million students took the NEET exam, a number that has been rising annually by 10%-16.5%, highlighting the exam's widespread popularity and importance. Among the 1.15 million candidates who qualified in 2023, only two scored full marks (720/720), one scored 716/720, seventeen scored 715/720, and six scored 711/720 [41]. NEET's rigorous nature,

coupled with its widespread adoption, underscores its importance as the primary evaluation tool for determining students' knowledge, aptitude, and readiness for pursuing medical and dental education at the undergraduate level [42].

In this investigation, to evaluate the performance of the three mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET 2023 exam, we employed rigorous statistical analyses. We scrutinized each model's performance across three pivotal frameworks: overall comparison, subject-level comparison, and topic-level comparison. The outcomes of our study can help pre-med students making informed decisions about incorporating LLMs into their test preparation strategies. To the best of our knowledge, this marks the first endeavor to undertake such a study.

Methodology

Question set selection and preparation

In this paper, we test the performance of the three LLMs on NEET-2023, which was obtained as a portable document file. Although the exam consists of 200 questions, due to the presence of illustrations and diagrams, it was not possible to process all the questions. As a result, we excluded questions with illustrations, resulting in a set of 175 questions for our study. This sample size is large enough to statistically justify each model's performance on the entire exam, with a 95% confidence level and a 5% margin of error. The selected questions were then manually presented to Bard, GPT-3.5, and GPT-4, and the responses were documented in Microsoft Excel.

Data Analysis

1) Prediction Performance

We compared responses generated by each model against the correct answers from the official answer key from the NEET website (<https://neet.nta.nic.in/>). Based on this comparison, the responses were either marked as correct (1) or incorrect (0). Excel's built-in functionalities were then utilized to generate four comparison metrics. In particular, we calculated accuracy and precision defined as follows.

- 1) Accuracy is defined as the percentage of correct responses obtained by a model. In the context of this research, accuracy was obtained using the following formula:

$$\text{Accuracy} = \text{Correct Responses} / \text{Total Responses}$$

- 2) Precision is defined as the number of true positive (TP) predictions by the sum of true positive (TP) and false positive (FP) predictions. In our context, a TP refers to the official correct response provided by NTA where the compared models agree; a FP refers to an incorrect response on which the compared models agree. The formula is given below:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = \text{Correct Responses} / \text{Total Consensus}$$

2) Scoring Performance

All the responses were compared against the official answer keys provided by NTA and the total score obtained by each chatbot was calculated. In accordance with the rules, each correct response

received 4 points and each incorrect response led to a 1-point deduction. We combined Zoology and Botany into one category, Biology, because the topic-level categorization included questions from both subjects. The total points scored by each model were then divided by maximum possible points, i.e., 720 to obtain the overall score percentage. To obtain subject-level percentages, we divided total points obtained by each model by the maximum points that could be obtained in that subject. Similarly, topic-level percentages were calculated by dividing the total points obtained by each model by the maximum possible points within that topic, which varied from topic to topic.

Results

We found that GPT-4 demonstrates higher accuracy and precision compared to GPT-3.5 and Bard. It also consistently outperforms the other models across subjects and topics. GPT-3.5 and Bard show variations in their performances, with specific strengths in certain subjects and topics.

1) Prediction Performance

a) **Accuracy.** The accuracy rates of the models are as follows:

- GPT-4 achieves the highest accuracy rate of approximately 54.29% by correctly identifying 95 responses.
- GPT-3.5 demonstrates an accuracy of 36.57% with 64 correct responses.
- Bard achieves the lowest accuracy of approximately 33.14% based on 58 correct answers.

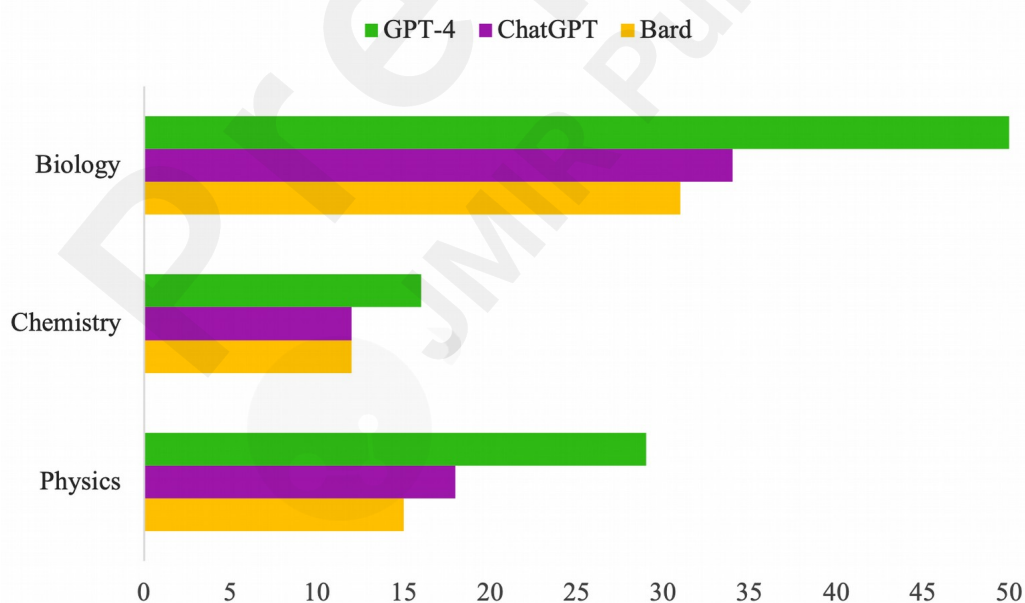


Figure 1. Number of correct responses in each subject per model.

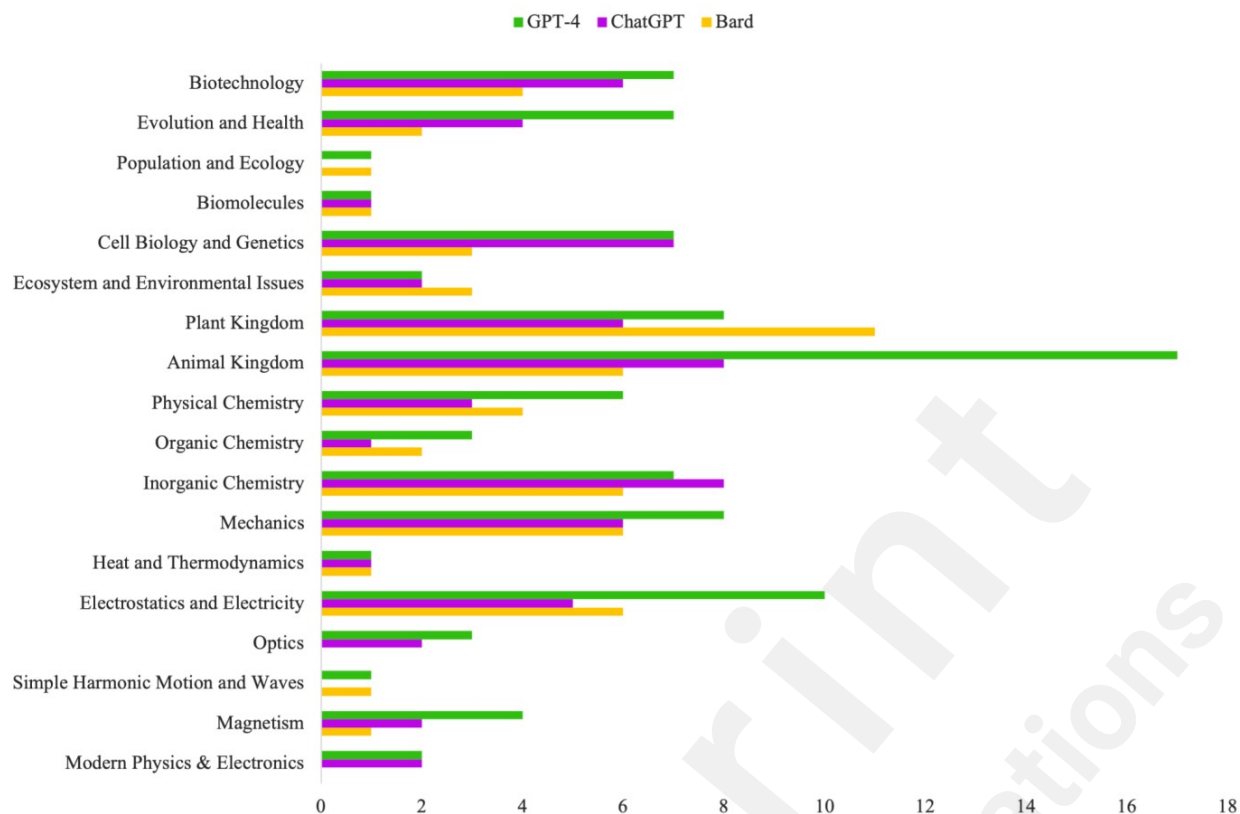


Figure 2. Number of correct responses for each topic per model.

Figure 1 presents the number of correct responses obtained by each model in each of the three subject areas covered by NEET. It is evident that GPT-4 is consistently more accurate than both Bard and GPT-3.5 in all three subjects. For each subject, the number of correct responses obtained by GPT-3.5 and Bard differs by ± 3 , indicating relatively similar subject-level accuracy rates. On the other hand, GPT-4 is significantly more accurate than the other models, generating 4 to 16 more correct answers per subject. In Physics, GPT-4 achieves 72.5% accuracy, followed by GPT-3.5 with 45%, and Bard with 37.5%. Similarly, in Chemistry, GPT-4's accuracy rate is 44.44%, while GPT-3.5 and Bard achieve an accuracy rate of 33.33%. Shifting to Biology, GPT-4 maintains its lead with 50.5% accuracy, followed by GPT-3.5 with 34.34%, and then Bard with 31.31%.

Figure 2 displays the number of correct responses obtained from each model in various topics. GPT-4 is the most accurate in 9 (ranging from 90.90% to 33.33%) out of 18 topics. Moreover, for at least half (2-4) of the topics in each subject, GPT-4 demonstrated the highest accuracy. GPT-3.5 is the most accurate (53.33%) in Inorganic Chemistry. In addition, it is more accurate (66.66% to 33.33%) than Bard in 7 topics across the three subjects. However, it has zero accuracy in Population & Ecology (Biology) and Simple Harmonic Motion & Waves (Physics). Bard is the most accurate in Plant Kingdom (44%) and Ecosystem & Environment Issues (60%). Furthermore, it is more accurate (100% to 33.33%) than GPT-3.5 in 5 topics across all three subjects. However, it has zero accuracy for two physics topics, namely Modern Physics & Electronics and Optics. GPT-4 and GPT-3.5 have similar accuracies in 1 physics topic (Modern Physics & Electronics (50%)) and 2 biology topics (Cell Biology & Genetics (43.75%) and Ecosystem & Environmental Issues (40%)). GPT-4 and Bard are 100% accurate in Simple Harmonic Motion & Waves. All three models are at the same level of accuracy in Biomolecules (33.33%) and Heat & Thermodynamics (33.33%).

In a nutshell, GPT-4 has a higher accuracy across a wide range of topics (15 out of 18), while GPT-3.5's and Bard's accuracies are well below GPT-4's. Moreover, they show variations in their accuracies across topics.

b) **Precision.** The precision values for the pairs are approximately:

- i) Bard and GPT-3.5 agree with each other on 69 responses, of which 32 are correct, giving the pair a precision of 0.4638 or 46.38%
- ii) Bard and GPT-4 match on 75 responses, of which 39 are correct. This results in a precision of 0.5200 or 52.00%
- iii) GPT-3.5 and GPT-4 concur on 79 responses, of which 44 are correct. This gives the pair a precision value of 0.5569 or 55.69%
- iv) All three models agree amongst each other on 49 responses, of which 24 are correct. The precision value of the trio is approximately 0.4898 or 48.98%.

2) Overall Scoring Performance

GPT-4 achieved the highest score with 300 marks (43%), outperforming GPT-3.5, which scored 145 marks (21%), and Bard, which obtained 115 marks (16%). To qualify for the NEET-2023 entrance test, candidates needed to secure at least 137 marks out of 720, which represents 19% of the total marks. It is evident that GPT-4 passed the entrance test with flying colors, showcasing exceptional performance. On the other hand, GPT-3.5 managed to qualify but with a significantly lower score. However, Bard failed to meet the qualifying criteria and, hence, did not pass the test.

a) Subject-wise Scoring

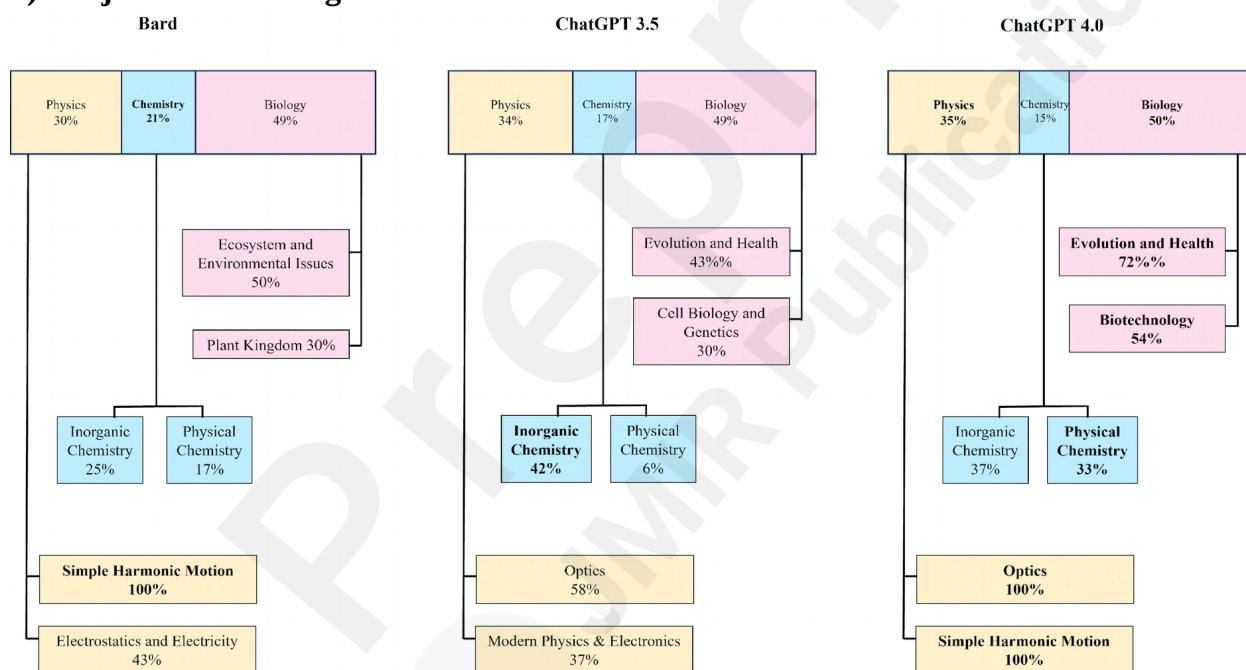


Fig. 3: A

comparison of the subject-wise distribution of Bard, GPT-3.5 and GPT-4 (a) Biology (b) Chemistry (c) Physics.

The figure 3 displays the subject-wise score percentages obtained by Bard, GPT-3.5, and GPT-4. Notably, all three chatbots obtained 49-50% of their total scores exclusively from the biology category. GPT-4 secured 35% and 15% of the total marks in physics and chemistry, respectively. GPT-3.5 received 34% and 17% of the total marks in physics and chemistry. Bard obtained 30% and 21% of the total marks from physics and chemistry, respectively. GPT-4 demonstrates superior performance in Biology and Physics, while Bard excels in Chemistry overall.

b) Topic-wise Scoring

In Table 1, marks obtained by GPT-3.5, GPT-4, and Bard across different subjects are presented. Remarkably, GPT-4 achieved the highest overall score of 43%, surpassing GPT-3.5 (21%) and Bard (16%). In the subject of Physics, GPT-4 performed exceptionally well in all the topics. However, in the Chemistry subject, it lagged behind in Inorganic Chemistry. When comparing the GPT-3.5 models, Bard obtained the highest marks in Plant Kingdom and Ecosystem and Environmental Issues

exclusively. Both versions of GPT-3.5 attained the same highest scores in Cell Biology and Genetics (Biology) and Modern Physics & Electronics (Physics). Nonetheless, GPT-3.5 demonstrated excellent performance in Inorganic Chemistry.

Table 1: Subject and topic level marks for Bard, GPT-3.5 and GPT-4. Green denotes the top scorer and red the lowest scorer within the category. Rest of the scores are colored yellow.

	Marks obtained		
	Bard	GPT-3.5	GPT-4
Overall	115 (16%)	145 (21%)	300 (43%)
Biology	56	71	151
Animal Kingdom	6	16	61
Plant Kingdom	30	5	15
Ecosystem and Environmental Issues	10	5	5
Cell Biology and Genetics	-1	19	19
Biomolecules	2	2	2
Population and Ecology	-1	-6	-1
Evolution and Health	1	11	26
Biotechnology	9	19	24
Chemistry	24	24	44
Inorganic Chemistry	15	25	20
Organic Chemistry	1	-4	6
Physical Chemistry	8	3	18
Physics	35	50	105
Modern Physics and Electronics	-4	6	6
Magnetism	-1	4	14
Simple Harmonic Motion and Waves	4	-1	4
Optics	-3	7	12
Electrostatics and Electricity	19	14	39
Heat and Thermodynamics	2	2	2
Mechanics	18	18	28

Discussion:

We evaluated the decision-making performance of three models—Bard, GPT, and GPT-4—using accuracy, precision, and test scores for the NEET-2023 entrance test. Subject-wise and topic-wise analyses were also conducted. GPT-4 consistently outperformed Bard and GPT across all subjects, achieving the highest accuracy rates: 72.5% in Physics, 44.44% in Chemistry, and 50.5% in Biology. Topic-wise comparisons also demonstrated GPT-4's excellence in 15 out of 19 topics, with Bard and GPT excelling in certain topics. Particularly, Bard excelled in Simple Harmonic Motion and Waves, while GPT showed strength in Inorganic Chemistry. Overall, GPT-4 emerged as the top performer, excelling in both subjects and specific topics. Our findings are in line with previous studies that have also examined how LLMs perform on exams related to medical education. [29] found that GPT-3.5 performs at or above the median performance of the MCAT takers. [33] reported that GPT-4

outperformed both GPT-3.5 and Bard by achieving the highest score of 82.6% in specialized questions prepared for neurosurgery oral board examinations. [43] found that GPT-3.5 answered about two-thirds of the multiple-choice questions correctly and outperformed nearly all medical students in years 1-3 of their studies. [30] reported that GPT-3.5's performance on the USMLE was either at or near the minimum passing threshold, even without domain-specific fine-tuning. Below we present both practical and research implications of our findings to enrich the existing literature.

Implications

i) Practical Implications

The findings have important implications for users who need to select a model based on specific requirements and aspired score. The subject and topic level scores highlight the suitability of different models for different domains. GPT-4 appears to have the highest score (43%), followed by GPT-3.5 (21%), and then Bard (16%). This demonstrates that Bard was not able to pass the NEET 2023 admission exam and GPT-3.5 was only 2% marks away from the cutoff score, which is 19%. Although GPT-4's overall score is higher than Bard's and GPT-3.5's overall scores, it falls short in Chemistry, receiving the lowest scores (15%) of the three. In chemistry, Bard outperforms the others overall, however, GPT-3.5 aces in Inorganic Chemistry indicating it has a better understanding of the topic (Fig. 3 and Table 1).

Although GPT-4 appears to be the preferred choice for NEET preparation, it is important to note that GPT-4 is a subscription-based service and the pricing model is uniform across the globe, which makes this model less accessible to the general audience in some parts of the world, particularly low-income countries. When cost is an issue, prospective medical school students might consider using GPT-3.5 and Bard in tandem to develop specialized knowledge and expertise in specific subject topics. These results suggest that, in the absence of GPT-4, it would be beneficial to consider both GPT-3.5 and Bard together for topic-specific preparation of NEET. This indicates that students may find themselves confused about the correct responses while cross-checking answers with these models. Therefore, it is recommended that, for exam preparation, students do not solely rely on these models; instead, they should consult primary sources in conjunction with these models.

ii) Research Implications

While scoring performance comparisons help us evaluate whether these models are able to ace the NEET-2023 exam or not, prediction performance comparisons help us evaluate their long-term performance beyond NEET 2023. The models' predictive accuracy rates match their scoring performance. GPT-4 demonstrated the highest accuracy rate among the three models, indicating its superior capability to provide correct responses and its reliability as an accurate study partner. However, there is still plenty of room for improvement, since its accuracy is still at 54.29%, suggesting that anyone using this model for exam preparation will be exposed to little over 50% accurate information. GPT-3.5 (36.57%) and Bard (33.14%) have similar overall accuracy rates that are much lower compared to GPT-3.5's, suggesting that these two models require significant fine-tuning.

The subject and topic level accuracy comparisons highlight specific areas where these models could benefit from domain-specific enhancements. GPT-4 demonstrates superior accuracy across all three subjects and 15 topics but requires further improvements in 3 topics, i.e., Ecosystem & Environmental Issues, Plant Kingdom and Inorganic Chemistry. GPT-4 excels in at least one topic from each subject category, including Simple Harmonic Motion & Waves and Optics in Physics, Physical Chemistry in Chemistry, and Evolution and Biotechnology in Biology. Bard outshines in Simple Harmonic Motion & Waves, and GPT-3.5 excels notably in Inorganic Chemistry. GPT-3.5, besides requiring improvements in its overall prediction capabilities, needs to develop predictive expertise in Population & Ecology (Biology) and Simple Harmonic Motion & Waves (Physics). Similarly, Bard needs to develop predictive capabilities in Modern Physics & Electronics and Optics,

in addition to requiring substantial enhancements in its overall predictive capabilities.

In summary, the implications and applications of this study on LLM and education are far-reaching. Firstly, it could serve as a benchmark for evaluating and improving LLMs' performance in exams and other educational tasks, enhancing the overall effectiveness of these models in educational settings. Secondly, the utilization of LLMs as tutors, mentors, or peers has the potential to significantly enhance students' learning outcomes and motivation, particularly in a country like India with a vast student population and diverse learning needs. Lastly, this approach could serve as a platform to explore and address ethical and social concerns related to LLMs in education, such as issues of fairness, bias, privacy, and accountability, ensuring responsible and informed use of these models in educational contexts.

Limitations and further research

Like any other research, our study has certain limitations that should be considered carefully. It is important to note that our study does not involve direct input from actual students, teachers, or medical school boards to understand their perspectives on these mainstream LLMs' capability to answer questions on basic science concepts. Moreover, we do not know how prospective examinees are using these models for exam preparation and whether they trust them for critical issues such as exam preparation.

LLMs have evolved considerably just in the last six months. Therefore, it is clear that the results of this study will have to be revisited at a later stage. For example, it is possible (and likely) that the relative performance of the different models will change over time. While Bard is currently lagging behind GPT-3.5 in this area, improvements to the model could mean that it might catch up to GPT-3.5 in the future. Since there is currently an "AI race" among many technology firms [47-49] it is only a matter of time before new models are introduced that could perform better on these types of questions.

Conclusion:

In this study, we conducted a comparative analysis of three notable chatbots, Bard, GPT-3.5, and GPT-4 to evaluate their performance on NEET 2023, a highly competitive medical school entrance examination in India. The study involved the preparation of NEET 2023 questions for the chatbots, data collection, data analysis, and scoring performance assessments.

Our results indicate that GPT-4 not only passed the NEET-2023 entrance test with 43% of marks but also demonstrated higher accuracy and precision compared to both GPT-3.5 and Bard. Particularly, GPT-4 consistently outperformed the other models across subjects and topics, achieving an accuracy rate of approximately 54.29%. GPT-3.5 and Bard, on the other hand, showed variations in their performances, with specific strengths in certain subjects and topics. Regarding subject-wise scoring, GPT-4 excelled in Physics and Biology while Bard performed well in Chemistry.

These findings shed light on the proficiency of LLMs in answering high-stakes examination questions, particularly in the context of medical entrance exams like NEET. GPT-4's superior performance and accuracy suggest its potential utility as a valuable resource for medical students seeking assistance in test preparation and knowledge acquisition. However, it is essential to note that despite their impressive performance, LLMs like Bard, GPT-3.5 and GPT-4 can sometimes generate text containing fabricated or incorrect information. This raises concerns about the credibility of information produced by LLMs, especially in educational settings where accuracy is crucial.

It is also important to acknowledge that LLMs, including GPT, come with both positive and negative consequences [44, 45]. Friederichs et al. [43] argue that the ability to acquire knowledge is a basic determinant of a physician's performance, and GPT-3.5 should be looked upon as a tool that provides easy access to a lot of relevant information, eventually aiding in clinical decision-making processes. On the other hand, authors in [46] have commented that GPT-3.5's success on exams like USMLE demonstrates the flaws of medical education, which is "mostly focused on the rote memorization of mechanistic models of health and disease" and does not reward critical thinking to the same extent.

Further research and development are warranted to address the limitations and challenges posed by

LLMs to ensure their reliable and accurate utilization in education and other domains. Moreover, future investigations can explore the suitability of LLMs in addressing the needs of diverse professional fields beyond medical entrance exams.

In conclusion, this study contributes valuable insights into the capabilities of Bard, GPT-3.5 and GPT-4 in handling high-stakes examination questions. As LLMs continue to evolve, their potential to revolutionize education and other industries remains promising, albeit with the need for continuous improvements and validation of their accuracy and reliability.

Funding statement: Not applicable.

Data availability: Data can be provided by an adequate request to the corresponding author.

Conflict of interest: We declare that there is no conflict of interest among the authors.

Author's contributions: Conceptualization: FF and SSS, Data acquisition: FF and DØM, Data analysis: FF and BMC, Writing and Drafting: FF, BMC, MN and SSS, Reviewing and proofreading: MN and DØM, Supervision: SSS.

References:

1. Farhat, F., Sohail, S. S., & Madsen, D. Ø. (2023). How trustworthy is ChatGPT? The case of bibliometric analyses. *Cogent Engineering*, 10(1), 2222988
2. Dwivedi, Y. K., Pandey, N., Currie, W., & Micu, A. (2023). Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. *International Journal of Contemporary Hospitality Management*.
3. Sohail, S. S., Madsen, D. Ø., Farhat, F., & Alam, M. A. (2023). ChatGPT and Vaccines: Can AI Chatbots Boost Awareness and Uptake?. *Annals of Biomedical Engineering*, 1-5.
4. Kovačević, D. (2023, March). Use of ChatGPT in ESP teaching process. In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE
5. Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1), 15
6. Beerbaum, D. O. (2023). Generative Artificial Intelligence (GAI) with Chat GPT for Accounting—a business case. Available at SSRN 4385651.
7. Silva, M., & Costa, D. (2022). Chat GPT and Human Resource Management. In *Conferência-Investigação e Intervenção em Recursos Humanos* (No. 11).
8. McGee, R. W. (2023). What Are the Top 20 Questions in Sociology? A ChatGPT Reply. Working Paper, April 8, 2023. <https://ssrn.com/abstract=4413441>.
9. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to ChatGPT. *arXiv preprint arXiv:2303.04226*.
10. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
11. Aljanabi, M., Ghazi, M., Ali, A. H., Abed, S. A. & ChatGPT (2023), 'ChatGPT: Open possibilities', *Iraqi Journal for Computer Science and Mathematics* 4(1), 62 – 64.
12. Ray, P. P. (2023), 'ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope', *Internet of Things and Cyber-Physical Systems*.
13. Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.
14. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
15. Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
16. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A. & Fiedel, N. (2022).

- Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
- 17.OpenAI (2022), 'ChatGPT', <https://openai.com/ChatGPT> [Online; accessed 29-March-2023].
- 18.Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T & Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.
- 19.Borji, A., & Mohammadian, M. (2023). Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. GPT-4, Claude, and Bard (June 12, 2023).
- 20.Tlili, A., Shehata, B., Adarkwah, M.A. et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* 10, 15 (2023). <https://doi.org/10.1186/s40561-023-00237-x>
- 21.Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research. *Journal of Educational Technology and Innovation*, 5(1).
- 22.Abd-alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P.M., Latifi S., Aziz, S., Damseh, R., AlabedAlrazak, S., Sheikh, J. (2023). Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ* 2023;9:e48291
- 23.Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*, nov 2022. doi: 10.1145/3571730.
- 24.Newton, P. M. &Xiomeriti, M. (2023). ChatGPT Performance on MCQ Exams in Higher Education. A Pragmatic Scoping Review. edarxiv: Available online: <https://edarxiv.org/sytu3>
- 25.Terwiesch, C. (2023). Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania. Available online: <https://mackinstitute.wharton.upenn.edu/wpcontent/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf>
- 26.Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., ... & Kuruppu, C. (2023). The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?. *Issues in Accounting Education*, 1-28.
- 27.Choi, Jonathan H. and Hickman, Kristin E. and Monahan, Amy and Schwarcz, Daniel B., ChatGPT Goes to Law School (January 23, 2023). *Journal of Legal Education* (Forthcoming), Available at SSRN: <https://ssrn.com/abstract=4335905> or <http://dx.doi.org/10.2139/ssrn.4335905>
- 28.Zhu, L., Mou, W., Yang, T., & Chen, R. (2023). ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. *Resuscitation*, 188, 109783.
- 30.Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198.
- 31.Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., &Chartash, D. (2023). 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*, 9(1), e45312.
- 32.Teebagy, S., Colwell, L., Wood, E., Yaghy, A., & Faustina, M. (2023). Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. *medRxiv*, 2023-04.
- 33.Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z. & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, 10-1227.
- 35.Liu, X., Fang, C., & Wang, J. (2023). Performance of ChatGPT on Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese. *medRxiv*, 2023-04.
- 38.Arumugam, V., & Mamilla, R. (2019). NEET for medics: a guarantee of quality? An exploratory study. *Quality Assurance in Education*, 27(2), 197-222.
- 41.NEET Result 2023: Two students achieve perfect score of 720, national cut-offs increase across categories | Mint #AskBetterQuestions (livemint.com)

- 42.Gohain, M.P. (2023). Over 20 lakh applicants for NEET UG, Times of India, April 20, 2023 (accessed July 5, 2023). <https://timesofindia.indiatimes.com/education/news/over-20-lakh-applicants-for-neet-ug-2023/articleshow/99621365.cms>
- 43.Friederichs, H., Friederichs, W. J., & März, M. (2023). ChatGPT in medical school: how successful is AI in progress testing?. *Medical Education Online*, 28(1), 2220920.
- 44.Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., ... & Mansoor, W. (2023). Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions. *arXiv preprint arXiv:2307.14107*.
- 45.Sohail, S. S. (2023). A Promising Start and Not a Panacea: ChatGPT's Early Impact and Potential in Medical Science and Biomedical Engineering Research, In *Annals of Biomedical Engineering* [in press].
- 46.Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digital Health*, 2(2), e0000205.
- 47.Rahaman, Md. Saidur and Ahsan, M. M. Tahmid and Anjum, Nishath and Rahman, Md. Mizanur and Rahman, Md Nafizur, The AI Race is on! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article (February 8, 2023). Available at SSRN: <https://ssrn.com/abstract=4351785> or <http://dx.doi.org/10.2139/ssrn.4351785>.
48. Farhat, F. (2023). ChatGPT as a Complementary Mental Health Resource: A Boon or a Bane. *Annals of Biomedical Engineering*, 1-4.
49. Sohail SS, Madsen DØ, Himeur Y and Ashraf M (2023) Using ChatGPT to navigate ambivalent and contradictory research findings on artificial intelligence. *Front. Artif. Intell.* 6:1195797. doi: 10.3389/frai.2023.1195797