Article

Not peer-reviewed version

Beyond Human Understanding: Benchmarking Language Models for Polish Cariology Expertise

Simona Wojcik , Anna Rulkiewicz , Piotr Pruszczyk , Wojciech Lisik , Marcin Poboży , Iwona Pilchowska , Justyna Domienik-Karlowicz *

Posted Date: 18 September 2023

doi: 10.20944/preprints202309.1100.v1

Keywords: ChatGPT; google bard; innovations; AI in medicine; health IT; artificial intelligence; large language model; medical education; language processing; virtual teaching assistant

Article

Beyond Human Understanding: Benchmarking Language Models for Polish Cariology Expertise

Simona Wójcik [1], Anna Rulkiewicz [1], Piotr Pruszczyk [2], Wojciech Lisik [3], Marcin Pobozy [4], Iwona Pilchowska [1] and Justyna Domienik-Karłowicz [1,2,*]

[1] LUX MED
[2] Department of Internal Medicine and Cardiology with The Center for Diagnosis and Treatment of Thromboembolism, Medical University of Warsaw, Warsaw, Poland
[3] Department of General and Transplantation Surgery, Medical University of Warsaw, Warsaw, Poland
[4] NZOZ Cichowski Pobozy, Maciejowice, Poland
* Correspondence: justyna.domienik@wum.edu.pl

Abstract: The growing dependence on large language models (LLM)s highlights the urgent need to deepen trust in these technologies. Regular, rigorous validation of their expertise, especially in nuanced and intricate scenarios, is essential to ensure their readiness for clinical applications. Our study pioneers the exploration of LLM utility in the field of cardiology. We stand at the cusp of a transformative era where mature AI and LLMs, notably ChatGPT, GPT-4, and Google Bard, are poised to influence healthcare significantly. Recently, we put three available LLMs, OpenAI's ChatGPT-3.5, GPT-4.0, and Google's Bard, to the test against a significant Polish medical specialization licensing exam (PES). The exams cover the scope of completed specialist training, focusing on diagnostic and therapeutic procedures, excluding invasive medical procedures and interventions. In our analysis, GPT-4 consistently outperformed the others, ranking first, with and Google Bard and ChatGPT-3.5 following, respectively. The performance metrics underscore GPT-4's notable potential in medical applications. Given a score improvement of over 23.5 % between two AI models released just four months apart, clinicians must stay informed and up-to-date about these rapidly evolving tools and their potential applications to clinical practice. Our results provide a snapshot of the current capabilities of these models, highlighting the nuanced performance differences when confronted with identical questions.

Keywords: ChatGPT; google bard; innovations; AI in medicine; health IT; artificial intelligence; large language model; medical education; language processing; virtual teaching assistant

1. Introduction

The competition to implement artificial intelligence (AI) technology in impactful ways has never been more intense. In particular, generative AI has emerged as a dominant force, ushering in a realm of applications and potential value. The spotlight has recently been on Generative Pre-trained Transformer (GPT) models, gaining prominence with the introduction of OpenAI's ChatGPT, a technology often hailed as a game-changer [1]. GPT technology leverages vast amounts of publicly available digital content data (in the field of natural language processing) to process and generate text that closely mimics human language, showcasing creativity in producing convincing written content on a wide array of subjects. The development of ChatGPT has undergone multiple phases. While ChatGPT became accessible to the public in November 2022 under the label "GPT-3.5," OpenAI launched an upgraded model, GPT-4, on March 14, 2023. Similar to its precursor, GPT-4 underwent training using a combination of supervised and unsupervised learning methodologies on an extensive dataset of internet text, followed by refinement through reinforcement learning using human feedback. Notably, GPT-3 boasts an astonishing 175 billion parameters, a tenfold increase compared to any previously developed language model. GPT-3 serves as the foundational NLP engine underlying the recently developed language model, ChatGPT, which has captured the interest of diverse domains, including but not limited to education and healthcare. Since its launch on November 30, 2022, ChatGPT rapidly garnered over one million subscribers within a week [2]. More recently, an even more advanced model, GPT-4, was unveiled on March 14, boasting an astounding 170 trillion parameters, signifying a remarkable leap in computational processing capability compared to its predecessor [3].

1. Medical Milestones: Assessing ChatGPT's Proficiency in Healthcare Examinations

Since its public introduction, ChatGPT and its potential applications have been the focus of extensive research. Notably, many in the educational community have heralded it as a transformative tool that might reshape traditional assignments and assessments. This sentiment is not unfounded; there is a growing body of evidence suggesting ChatGPT's adeptness in passing reputable exams such as the USMLE [6–8,10]. Such achievements underscore the need to reconsider and adapt current assessment tools within healthcare education.

Previous research has indicated that ChatGPT possesses the capability to address higher-order questions, even those as challenging as the USMLE, showcasing its logical and reasoning capacities [4–8]. A notable study in ophthalmology by Antaki et al. revealed that ChatGPT's performance was on par with an average first-year resident [9]. Furthermore, when evaluated on the Plastic Surgery In-Service Examination, its performance rivaled that of plastic surgery residents at various stages of their training. Specifically, of the 1129 questions assessed, ChatGPT answered 630 (55.8%) correctly. Intriguingly, its peak performance was observed in the 2021 exam (60.1%) and the comprehensive section (58.7%). When juxtaposed with plastic surgery residents in 2022, ChatGPT would place in the 49th percentile for first-year residents, 13th percentile for second-year residents, and significantly lower for subsequent years [10]. Further testament to ChatGPT's capabilities is its success with the rigorous European Exam in Core Cardiology (EECC) — a decisive assessment for cardiology specialty training completion across several countries. The EECC rigorously tests a trainee's knowledge spanning pathophysiology, clinical reasoning, and guideline-endorsed medical management through 120 multiple-choice questions (MCQs). Although the pass rate fluctuates, it hovers around 60%. Remarkably, ChatGPT's performance consistently surpassed this benchmark in most evaluations [11]. These findings underscore the burgeoning role of language models like ChatGPT in healthcare.

2. Rise of the Competitors: Google's Bard Challenges ChatGPT's Dominance in the AI Landscape

In the wake of the ascent of ChatGPT and GPT-4, numerous premier software entities have unveiled their own AI language models, epitomizing the progressive strides in AI. A notable entrant is Bard, an innovation by Alphabet Inc., Google's parent entity. Introduced on March 21, 2023, Bard has been pivotal in redefining the chatbot landscape, heralding fresh discourse about the trajectory of search technologies. In today's dynamic AI domain, Bard, powered by Google's LaMDA and OpenAI's ChatGPT, stands as vanguards, representing potent forces in the race for market preeminence, with both expected to undertake similar digital roles. However, amidst their triumphs, they are emblematic of the eternal tug-of-war seen in the tech cosmos. This cyclic competition is reminiscent of past tech face-offs, such as Samsung versus Apple or iOS versus Android [12], elucidating the age-old axiom that perpetual success is rooted in adaptability and prompt innovation assimilation. This adage seems apt, with the current epoch set for a pivotal face-off between these AI-driven conversational giants [13].

A salient differentiator for Bard vis-à-vis ChatGPT and GPT-4 is its prowess in real-time web information assimilation during response generation. In contrast, ChatGPT and GPT-4 are anchored to pre-existing knowledge up to September 2021, devoid of present web crawling competencies. This inherent capacity in Bard might equip users with more contemporaneous and contextually relevant data.

Recent evaluations accentuate the role of these tools in empowering users with informed choices, underscoring the multifaceted nature of AI content platforms. While ChatGPT showcased lesser internet dependency, hinting at more remarkable originality, Bard manifested a broader internet source spectrum. Such insights underscore the need for a nuanced assessment of these platforms, considering their intrinsic attributes, especially concerning content authenticity and originality. Both ChatGPT and Bard have made considerable strides in AI content generation. Nevertheless, they are not without limitations, warranting continuous refinements for optimized performance. With AI's relentless evolution, addressing these subtleties will inevitably augment their utility across diverse use cases [14].

Both Google and OpenAI proactively acknowledge potential inaccuracies or biases in their chatbot outputs and advocate for user discretion. Google's proactive redressal is manifested in Bard's "drafts" feature, offering a gamut of responses, while ChatGPT, though defaulting to a singular output, is pliable to generate alternates upon user solicitation.

Response accuracy in chatbot refers to the percentage of correct responses to user queries or inputs [60]. Both Google and OpenAI acknowledge the possibility of their chatbots providing inaccurate or biased information and recommend users to verify responses [61,62]. Google's approach to addressing limitations is evident in Bard, where users are presented with multiple response options, called" drafts," allowing for exploring and selecting the most resonant answer [63]. In contrast, ChatGPT typically provides a single response by default, although it can generate various versions upon request. Response accuracy in chatbot refers to the percentage of correct responses to user queries or inputs [60]. Both Google and OpenAI acknowledge the possibility of their chatbots providing inaccurate or biased information and recommend that users verify responses [61,62]. Google's approach to addressing limitations is evident in Bard, where users are presented with multiple response options, called" drafts," allowing for exploring and selecting the most resonant answer [63]. In contrast, ChatGPT typically provides a single response by default, although it can generate various versions upon request.

In the constantly evolving landscape of artificial intelligence, there has been a noticeable surge in academic interest surrounding the applications of AI chatbots in the medical domain. Recent publications spotlight the Bard, Google's cutting-edge chatbot, being subjected to rigorous examinations in a medical setting, reminiscent of the probing tests that OpenAI's GPT faced just a few months prior. [15]. These studies not only illuminate the potential of AI in assisting healthcare but also draw attention to the competitive landscape of technological advancements in this sphere. Against this backdrop, Google's Bard was put to the test, answering questions from the 2022 ASPS (American Society of Plastic Surgeons) In-Service Examination—a standardized annual examination aimed at evaluating the expertise of plastic surgery residents and prepping them for forthcoming written and oral boards. Bard's performance was commendable, surpassing over half of the first-year integrated residents by ranking in the 74th percentile. This was especially notable when juxtaposed against OpenAI's ChatGPT performance for the same examination [10,15].

Emerging research now not only delves into the individual efficacy of AI models like ChatGPT and Google Bard but also offers direct comparisons, a shift from the earlier trend of singular evaluations. An exemplar study conducted by Koga S et al. [16] concentrated on the prowess of ChatGPT and Google Bard in prognosticating neuropathologic diagnoses by leveraging clinical summaries. Drawing from a repository of 25 cases of neurodegenerative disorders collated from Mayo Clinic's esteemed brain bank Clinico-Pathological Conferences, and these models were tasked to furnish pathologic diagnoses along with pertinent rationales. The outcomes were illuminating. ChatGPT-3.5, ChatGPT-4, and Google Bard posted accuracy scores of 32%, 52%, and 40%, respectively. However, when considering the entirety of the answers provided, the proper diagnosis was nestled within their responses in 76% of cases for ChatGPT-3.5 and Google Bard and an even more commendable 84% for ChatGPT-4. Such revelations herald the ascendant potential of AI apparatuses like ChatGPT in the nuanced realm of neuropathology and suggest their prospective utility in enriching deliberations during clinicopathological assemblies [16]. Recent research by Rahsepar AA et al. compared the effectiveness of ChatGPT and Google Bard in answering lung cancer-related questions based on the 2022 Lung Imaging Reporting and Data System (Lung-RADS) guidelines. The results showed that ChatGPT-3.5 correctly answered 70.8% of the questions, while Google Bard managed a correct response rate of 51.7%. [17]. In a study by Patil NS and his team, ChatGPT-4 and Google's Bard were assessed for their proficiency in answering radiology board examination practice questions. The findings highlighted ChatGPT's superior performance in radiology knowledge, with an accuracy rate of 87.11% compared to Bard's 70.44%. ChatGPT outperformed Bard in specific areas such as neuroradiology, general and physics, nuclear medicine, pediatric radiology, and ultrasound [18].

2. Materials and Methods

Recently, we put three available LLMs, OpenAI's ChatGPT-3.5, GPT-4.0, and Google's Bard, to the test against a significant Polish medical specialization licensing exam (PES). The exams cover the scope of completed specialist training, focusing on diagnostic and therapeutic procedures, excluding invasive medical procedures and interventions. The tests and test questions are developed and established by the Center for Medical Education (CEM) in consultation with the national consultant responsible for the respective field of medicine. When a consultant is unavailable for a specific area, a related field consultant or their representative is consulted separately for each lot of medicine and each examination session [19]. The entrance test in the PES consists of solving 120 questions with five answer options, of which only one is correct. The test portion of PES is considered passed with a positive result when the physician achieves at least 60% of the maximum possible score.

In the case of this study, LLMs were deployed to address the cardiology section of the examination (Spring 2023) [20]. The CEM has been publishing test questions along with the correct answers after they have been used on a given exam within seven days of conducting that exam since the spring 2023 session.

Based on the difficulty of individual questions, it was assessed that 20.2% were easy questions (difficulty up to 0.5). Moderately difficult questions accounted for 47.1% of all questions (difficulty from 0.51 to 0.79), while the remaining part consisted of difficult questions (difficulty from 0.80). In terms of discriminative power (DP), questions with low discriminative power (53.8%; up to 0.3) dominated the test. Questions with average discriminative power (from 0.3 to 0.49) accounted for 31.1% of the questions, while those with high discriminative power (above 0.5) - 15.1%. A significant relationship was observed between the discriminative power and test difficulty (chi2 (4) = 45.90; p <0.001). It was found that questions with low discriminative power were more often classified into the group of questions with low or high difficulty, while the opposite relationship was observed for questions with moderate discriminative power.

Overall, a strong correlation was observed between the discriminative power index and the RPBI index for the correct answer (r = 0.859; p <0.001). A much weaker association was observed between the RPBI index for the correct answer and question difficulty (r = 0.318; p <0.001). There was no evidence to suggest that the DP indices and question difficulty were significantly correlated (r = -0.025; p >0.05).

From August 25 to August 28, 2023, responses to these queries were generated using two versions of ChatGPT (version GPT-3.5 and GPT-4.0, OpenAI, California) and Google Bard (Google LLC, Alphabet Inc., California). ChatGPT 3.5 and Google Bard are publicly accessible at no charge, whereas ChatGPT-4.0 requires a paid subscription. The ChatGPT-4 model encompasses more parameters and computational power than its predecessor, ChatGPT-3.5.29. Consequently, it is conceivable that ChatGPT 4.0 could better manage more intricate queries and tasks. We incorporated ChatGPT 3.5 and ChatGPT 4.0 into our evaluation to validate this hypothesis. The version of LLMs used in this study was the most up-to-date model at the time of publication.

The 120 selected questions were then input into each LLM-Chatbot; each was input as a 'standalone' query. Across all LLM-Chatbots, the conversation was reset after each query input to minimize memory retention bias. For this study, individualized prompts for each question were not employed; instead, an initial investigation was conducted to identify the prompts that yielded the most favorable responses.

The models' performance was further benchmarked against average scores achieved by human participants to provide a reference point. Comparative analysis was done to understand the performance variance between the different AI models.

3. Results

In our analysis, GPT-4 consistently outperformed the others, ranking first, with and Google Bard and ChatGPT- 3.5 following, respectively. The performance metrics underscore GPT-4's notable potential in medical applications.

The results of the conducted analysis of variance confirmed the presence of differences in the scores obtained between the efficiency of individual LLM-Chatbots models (F(2,117) = 9.85; p <0.001; eta2 = .144). The highest score was achieved by GPT4 4.0 (66.4% correct answers) and it was significantly higher than the score obtained by ChatGPT 3.5 (42.9%; significance of differences at p<0.001 level). No significant differences were observed between GPT4 4.0 and BARD (57.1%; p = 0.328). The difference between BARD and ChatGPT 3.5 was statistically significant (p=0.027). The obtained result is graphically presented below (Figure 1).

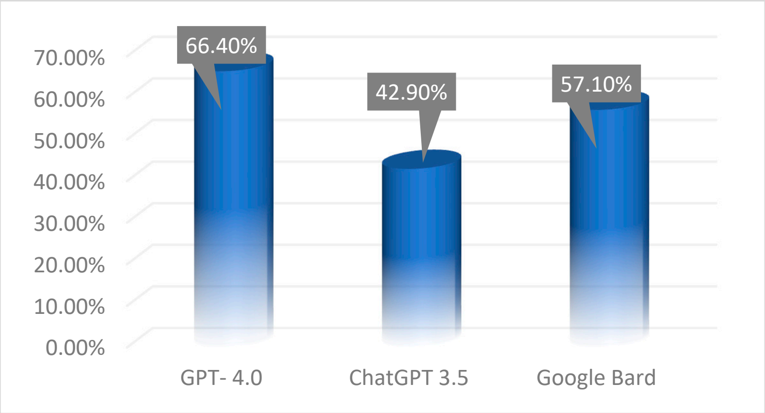

Figure 1. Average percentage of correct answers depending on the LLM source.

The chi-square analysis did not confirm the existence of a relationship between the correctness of the answers given by GPT 3.5., GPT 4, and Google Bard and the difficulty level as well as the discriminatory power of the exam question (see Tables 1–6). The results may indicate that the mentioned AI models' training is broad-based and does not necessarily favor any specific difficulty or discriminatory level of questions.

Table 1. Relationship between discriminative power and accuracy of responses – Google BARD.

	Incorrect answer	Correct answer
Low discriminative power	51,0%	55,9%
Moderate discriminative power	33,3%	29,4%
High discriminative power	15,7%	14,7%

Table 2. Relationship between discriminative power and accuracy of responses – GPT 4.0.

	Incorrect answer	Correct answer
Low discriminative power	47%	57%
Moderate discriminative power	32,5%	30,4%
High discriminative power	20,0%	12,7%

Table 3. Relationship between discriminative power and accuracy of responses – ChatGPT 3.5.

	Incorrect answer	Correct answer
Low discriminative power	52,9%	54,9%
Moderate discriminative power	29,4%	33,3%
High discriminative power	17,6%	11,8%

Table 4. Relationship between the difficulty of exam questions and the accuracy of answers- Google BARD.

	Incorrect answer	Correct answer
Low difficulty	25,5%	16,2%
Average difficulty	51,0%	44,1%
High difficulty	23,5%	39,7%

Table 5. Relationship between the difficulty of exam questions and the accuracy of answers- GPT-4.0.

	Incorrect answer	Correct answer
Low difficulty	30,0%	15,2%
Average difficulty	45,0%	48,1%
High difficulty	25,0%	36,7%

Table 6. Relationship between the difficulty of exam questions and the accuracy of answers- ChatGPT 3.5.

	Incorrect answer	Correct answer
Low difficulty	25,0%	13,7%
Average difficulty	50,0%	43,1%
High difficulty	25,0%	43,1%

It was not shown that the difficulty of the exam question had an impact on the existence of differences between the results of individual sources (F(4,232) = 0.24; p >0.05). Therefore, there is no basis to conclude that the result of each of the analyzed sources depended on the difficulty of the exam question. Knowing the effectiveness of the model in answering questions of varying difficulty levels, it can be predicted that its effectiveness will be similar for new, previously unanalyzed questions, regardless of their difficulty. The obtained result is graphically presented below (Figure 2).

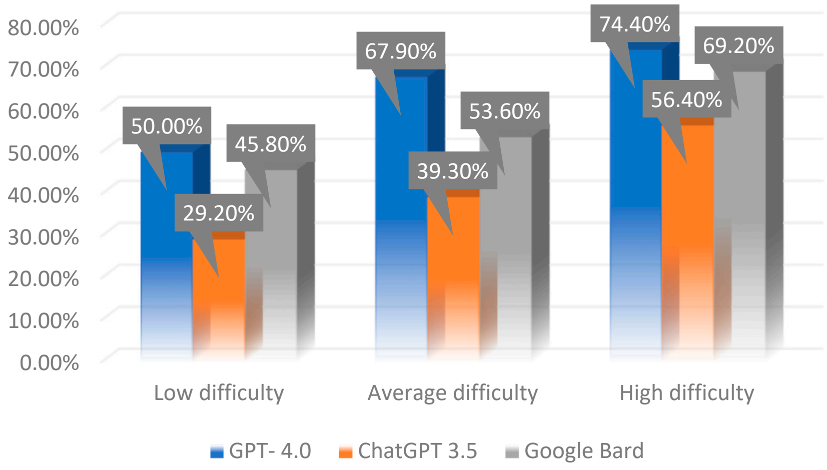

Figure 2. Average percentage of correct answers depending on the LLM source and difficulty of the exam question.

It was not demonstrated that the discriminative power of the question itself had an impact on the existence of differences between the results of individual LLMs (F(4,232) = 0.27; p >0.05). Therefore, there is no basis to conclude that the results of each of the analyzed LLMs depended on the discriminative power of the exam question. Since the effectiveness of the models was not related to the difficulty or discriminatory power of the questions, future research can focus on other factors

doi:10.20944/preprints202309.1100.v1

that influence the model's effectiveness in answering specialized questions. The obtained result is graphically presented below (Figure 3).

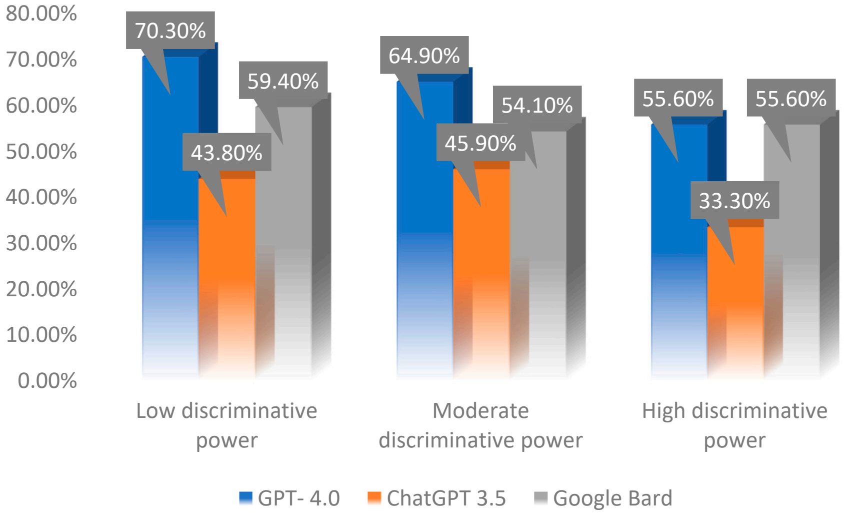

Figure 3. Average percentage of correct answers depending on the LLM source and discriminative power of the exam question.

The results of the Spearman's rho correlation analysis showed that the students' responses were significantly correlated only with ChatGPT 3.5 responses (rho = 0.203; p = 0.027). The correlation for student responses and Google BARD responses was at the level of a statistical trend (rho = 0.177; p = 0.054); the association in the case of GPT-4 proved to be statistically insignificant (rho = 0.126; p = 0.173). In addition, a significant relationship at the level of a statistical trend was observed for ChatGPT 3.5 when analyzing the relationship between the correct AI responses and the difficulty of the question according to student responses (ch^{i2}(2) = 5.91; p = 0.052). It was found that differences in the responses of students and ChatGPT 3.5 were visible in the case of the most difficult questions - then ChatGPT 3.5 answered significantly better. In the case of easy and medium-difficulty questions, no differences were observed. This could mean that for standard or less challenging content, both the model and the students have a similar understanding or approach. This also suggests that while students might struggle with challenging questions, ChatGPT 3.5 has the capacity to handle them relatively better.

In total, for 26.9% of the questions, every source provided the correct answer. Interestingly, for 16 questions, no source indicated the correct answer. It might be beneficial to further analyze the patterns of responses, particularly the incorrect ones, among the different LLM-Chatbot models. Understanding the nature of errors can shed light on the models' limitations or biases and suggest areas for improvement. Another interesting avenue could be integrating a feedback loop into the models, where their answers are validated and incorrect responses are corrected. Observing how quickly and effectively these models learn from feedback can be invaluable.

The obtained result is graphically presented below (Figure 4).

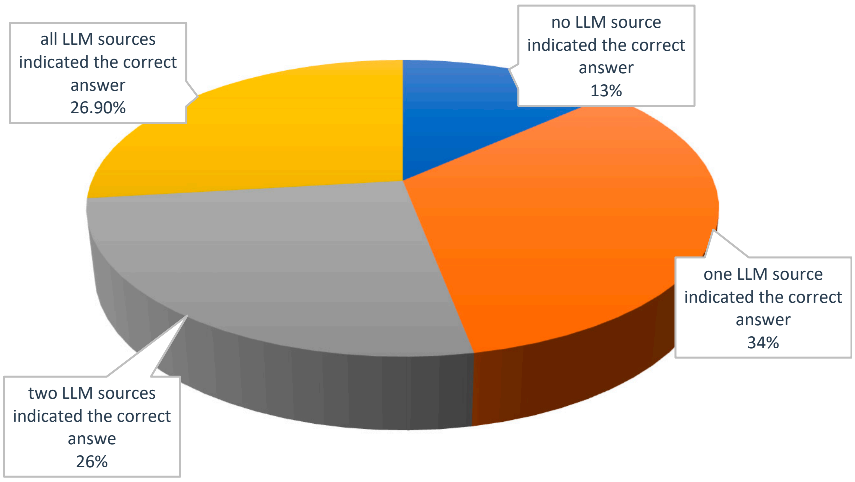

Figure 4. Percentage of correctly marked answers depending on the LLM source.

Correlation analysis indicated that there is a positive relationship between the difficulty of the question and the number of LLMs that indicated the correct answer (r = 0.260; p = 0.004). .No relationship was observed between the number of LLMs that indicated an answer and the discriminatory power of the exam question (p>0.05).

Nonetheless, no model demonstrated complete accuracy across all queries. This finding emphasizes the imperative for ongoing research and model optimization, which, if effectively addressed, can significantly enhance their utility in healthcare and cardiology.

GPT 4.0 displayed superior problem-solving prowess compared to the other evaluated Large Language Models (LLMs). Given a score improvement of over 23.5 % between two AI models released just four months apart, clinicians must stay informed and up-to-date about these rapidly evolving tools and their potential applications to clinical practice. Given the rapid evolution of LLMs, it is crucial to revisit and re-evaluate their capabilities periodically. Our results provide a snapshot of the current capabilities of these models, highlighting the nuanced performance differences when confronted with identical questions [21].

Our findings revealed that LLM-Chatbots, particularly ChatGPT-4.0, have the potential to deliver accurate and comprehensive responses to cardiology-related queries. This supports earlier studies by Ali et al. (2023) [22], and Raimondi et al. (2023), which found ChatGPT-4.0 outperformed other LLMs in neurosurgery and ophthalmology exams, respectively. This performance is in line with other previous research [22–24]. Since the study pointed out the potential of GPT-4 in medical applications, future research could explore how these models integrate and perform with medical databases or decision support systems.

4. Discussion

Interactions with LLMs have inherent limitations. We solely focused on evaluating the applicability of AI systems to resolve tests related to cardiology. Therefore, the findings and conclusions drawn from this study might not apply or be generalizable to other subjects or domains. Another study limitation is that different users might receive varied responses, mainly if they converse at different times. The way we phrase or rephrase questions can also influence the answers given by these models, potentially affecting our assessment of their capabilities [25]. Moreover the current findings, particularly regarding the lack of correlation between the AI models' performance and question difficulty or discriminatory power, highlight the need for future studies to explore other potential factors that might influence the models' accuracy in answering specialized questions. It is also important to remember that the version of Bard we tested is still in its early stages; its performance may enhance as its training data grows with time.

Testing AI on exams gives insights but does not guarantee how well it will perform in real medical situations. We have showcased LLMS capabilities in processing medical data and answering queries. Nevertheless, it should not replace doctors, especially considering their critical thinking, innovation, and creativity.

Additionally, our test prompts were in Polish. Other research indicates that LLMs often perform better with more straightforward questions and when the test language is English [26,27].

It would be valuable to expand the research to include other upcoming or less popular LLMs, to understand if there are any underdogs with niche specialties.

5. Conclusion

Our study pioneers the exploration of LLM utility in the field of cardiology at the time of publication.. We stand at the cusp of a transformative era where mature AI and LLMs, notably ChatGPT, GPT-4, and Google Bard, are poised to influence healthcare significantly. Evidence pointing to ChatGPT's success in medical licensing exams signals potential medical training and practice shifts. Medical institutions are encouraged to harness the power of AI, emphasizing the creation and validation of systems that deliver accurate and reliable information. By doing so, they can enrich teacher-student interactions, turning traditional lectures into engaging, collaborative learning sessions. Furthermore, if chatbots consistently meet or even exceed physician benchmarks in foundational exams, whether by showcasing proficient performance or evidencing deep knowledge, they could serve as supplementary tools in medical decision-making.

The growing dependence on LLMs highlights the urgent need to deepen trust in these technologies. Regular, rigorous validation of their expertise, especially in nuanced and intricate scenarios, is essential to ensure their readiness for clinical applications. Developing methodologies to quantify and understand 'hallucinations' or anomalies in their outputs is equally crucial. Only those LLMs adept at reducing and recognizing these irregularities should be considered for clinical integration. Our findings underscore the importance for medical professionals to stay updated on LLM advancements and carefully evaluate their potential roles in clinical environments.

References

1. Mathew, A. Is Artificial Intelligence a World Changer? A Case Study of OpenAI's Chat GPT. Recent Prog. Sci. Technol. 2023, 5, 35–42.
2. Rahimi, F.; Abadi, A.T.B. ChatGPT and publication ethics. Arch. Med. Res. 2023, 54, 272–274. [Google Scholar] [CrossRef] [PubMed]
3. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. 2023. Available online: https://www.techrxiv.org/articles/preprint/GPT-4_vs_GPT-3_5_A_Concise_Showdown/22312330 (accessed on April 15, 2023).
4. Kelly, S.M. (2023, January 26) ChatGPT passes exams from law and business schools. Retrieved from https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/
5. Terwiesch, C. (2023, January 17) Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. Retrieved from https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-newwhite-paper-by-christian-terwiesch
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198-e0000198
7. Antaki, F.; Touma, S.; Milad, D.; El-Khoury,J.; Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. medRxiv, 2023; Preprint.
8. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023 Jun;307(5):e230582. doi:10.1148/radiol.230582. Epub 2023 May 16. PMID: 37191485.
9. Antaki, F.; Touma, S.; Milad, D.; El-Khoury,J.; Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. medRxiv, 2023; Preprint.
10. Pooja Humar, BS and others, ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-service Examination, Aesthetic Surgery Journal, 2023; sjad130, Retrieved from: https://doi.org/10.1093/asj/sjad130

11. Corrigendum to: ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023 May 17;4(4):357. doi: 10.1093/ehjdh/ztad034. Erratum for: Eur Heart J Digit Health. 2023 Apr 24;4(3):279-281. PMID: 37538140; PMCID: PMC10393937.
12. B J Park, D Kim, Coopetition dynamics between giant entrants and incumbents in a new convergent segment: a case in the smartphone industry, Asian Journal of Technology Innovation, volume 29, issue 3, p. 455 - 476 Posted: 2021
13. 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
14. Ventayen, Randy Joy Magno, OpenAI ChatGPT, Google Bard, and Microsoft Bing: Similarity Index and Analysis of Artificial Intelligence-Based Contents (August 5, 2023). Available at SSRN: https://ssrn.com/abstract=4532471 or http://dx.doi.org/10.2139/ssrn.4532471
15. Daniel Najafali, BS and others, Bard Versus the 2022 American Society of Plastic Surgeons In-Service Examination: Performance on the examination in its Intern Year, Aesthetic Surgery Journal Open Forum, 2023; ojad066, https://doi.org/10.1093/asjof/ojad066
16. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. 2023 Aug 8:e13207. Doi: 10.1111/bpa.13207. Epub ahead of print. PMID: 37553205.
17. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023 Jun;307(5):e230922. doi: 10.1148/radiol.230922. PMID: 37310252.
18. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. Can Assoc Radiol J. 2023 Aug 14:8465371231193716. Doi: 10.1177/08465371231193716. Epub ahead of print. PMID: 37578849.
19. https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU19970280152/O/D19970152.pdf (Accessed August 2023)
20. https://cem.edu.pl/pytcem/wyswietl_pytania_pes.php (Accessed August 2023)
21. Kumari, Amita & Kumari, Anita & Singh, Amita & Singh, Sanjeet & Dhanvijay, Anup kumar & Pinjar, Mohammed Jaffer & Mondal, Himel & Juhi, Ayesha. (2023). Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus. 15. e43861. 10.7759/cureus.43861.
22. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Sullivan PLZ, et al. Performance of ChatGPT, 333 GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 334 2022:10.1227.
23. Morreel, Stefan & Verhoeven, Veronique & Mathysen, Danny. (2023). Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. 10.1101/2023.08.18.23294263.
24. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B, Cheng CY, Koh VTC, Tham YC. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023 August 23;95:104770. doi: 10.1016/j.ebiom.2023.104770. Epub ahead of print. PMID: 37625267
25. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023 Jun 26;15(6):e40977. doi: 10.7759/cureus.40977. PMID: 37519497; PMCID: PMC10372539.
26. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing 323 examination in Taiwan. J Chin Med Assoc. 2023;86(7):653-8. Epub 20230705. doi: 324 10.1097/jcma.0000000000000942. PubMed PMID: 37227901. 325
27. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style 326 Examination: Insights into Current Strengths and Limitations. Radiology. 2023;307(5):e230582. doi: 327 10.1148/radiol.230582.