sciendo

Original Study • DOI: 10.2478/ahem-2024-0006 • AHEM • Vol. 78 Issue 1/2024 • 111–116

AHEM

Advances in Hygiene and Experimental Medicine
Postępy Higieny i Medycyny Doświadczalnej

Performance of ChatGPT and GPT-4 on Polish National Specialty Exam (NSE) in Ophthalmology

Original Study

Marcin Ciekalski[1], Maciej Laskowski[1]*, Agnieszka Koperczak[1], Maria Śmierciak[1], Sebastian Sirek[2]

[1] Student Scientific Society, Department of Ophthalmology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Poland
[2] Department of Ophthalmology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Poland

Received 11 January, 2024, Accepted 19 June, 2024

Abstract

Introduction. Artificial intelligence (AI) has evolved significantly, driven by advancements in computing power and big data. Technologies like machine learning and deep learning have led to sophisticated models such as GPT-3.5 and GPT-4. This study assesses the performance of these AI models on the Polish National Specialty Exam in ophthalmology, exploring their potential to support research, education, and clinical decision-making in healthcare.
Materials and Methods. The study analyzed 98 questions from the Spring 2023 Polish NSE in Ophthalmology. Questions were categorized into five groups: Physiology & Diagnostics, Clinical & Case Questions, Treatment & Pharmacology, Surgery, and Pediatrics. GPT-3.5 and GPT-4 were tested for their accuracy in answering these questions, with a confidence rating from 1 to 5 assigned to each response. Statistical analyses, including the Chi-squared test and Mann-Whitney U test, were employed to compare the models' performance.
Results. GPT-4 demonstrated a significant improvement over GPT-3.5, correctly answering 63.3% of questions compared to GPT-3.5's 37.8%. GPT-4's performance met the passing criteria for the NSE. The models showed varying degrees of accuracy across different categories, with a notable gap in fields like surgery and pediatrics.
Conclusions. The study highlights the potential of GPT models in aiding clinical decisions and educational purposes in ophthalmology. However, it also underscores the models' limitations, particularly in specialized fields like surgery and pediatrics. The findings suggest that while AI models like GPT-3.5 and GPT-4 can significantly assist in the medical field, they require further development and fine-tuning to address specific challenges in various medical domains.

Keywords

ophthalmology • ChatGPT • Polish national specialty exam

1. Introduction

From its beginnings in the mid-20th century, artificial intelligence (AI) has experienced cycles of enthusiastic progress and quieter periods. Nevertheless, in recent years, the field has seen a revival in research and development, largely driven by the significant growth in computing capabilities and the emergence of big data. Technologies such as machine learning, neural networks, and deep learning aim to construct models capable of handling extensive datasets, recognizing patterns, and then producing insights or predictions, frequently outperforming human abilities in particular tasks [1].

The swift progression of artificial intelligence has shifted from fundamental research to practical, real-world applications. GPT-3.5, a model arising from this effort, demonstrated remarkable abilities based on its advanced architecture. Nonetheless, its successor, GPT-4, surpassed it, showing a 40% improvement in performance and showcasing much greater diversity in potential applications.

We can witness a noticeable change in clinical documentation procedures, with the support of AI tools becoming increasingly evident. By systematically analyzing extensive medical datasets, AI models can provide clinicians with evidence-based recommendations, potentially decreasing medical errors and enhancing the effectiveness of treatments.

As telemedicine becomes a vital component, especially in response to recent global health challenges, AI tools with real-time analytical capabilities can enhance patient care. Moreover, in the ever-evolving field of medical academia, where ongoing knowledge updates are crucial, GPT-3.5 and GPT-4 can serve as invaluable, continuously updated references. The incorporation of models like CLIP for the analysis of medical images further emphasizes the potential for enhanced diagnostic accuracy.

The domain of ophthalmology has experienced a considerable rise in research initiatives and the quantity of published research, particularly focusing on the application of AI tools in this branch of science [2]. In the field of ophthalmology, Deep Learning (DL) has been utilized for the analysis of fundus photographs, optical coherence tomography, and visual fields. It has demonstrated strong classification capabilities in identifying conditions such as diabetic retinopathy, retinopathy of prematurity, glaucoma-

* E-mail: maciek.lask@gmail.com

like disc, macular edema, and age-related macular degeneration [2]. In tasks like identifying and grading diabetic retinopathy AI systems have exhibited performance levels that are on par with or, in some cases, superior to experienced ophthalmologists. Nevertheless, despite these promising outcomes, only a limited number of AI systems have been put into practical use within clinical settings [3].

Employing questions sourced from the National Specialty Exam (NSE) is a viable approach for evaluating ChatGPT's proficiency in retrieving and analyzing specialized information within the ophthalmology field. Our research aimed to evaluate ChatGPT's performance in responding to queries and to analyze its strengths and weaknesses in comparison to human comprehension and reasoning.

2. Materials and methods

2.1. Examination and questions

The study was conducted within the time frame of October 19th to November 3rd, 2023. The research concentrated on a specific examination within the field of ophthalmology (Spring, 2023). This examination was selected randomly from the repository of available exams in the question archive database of the Medical Examinations Center in Lodz, Poland. Of the original set of 120 single-choice questions, each containing one correct response and four distractors, the Board of Examiners removed 21 questions due to their incompatibility with current knowledge standards. Consequently, a comprehensive analysis was performed on the remaining 98 questions.

The questions were categorized based on their content into five groups: Physiology & Diagnostics, Clinical & Case Questions, Treatment & Pharmacology, Surgery, Pediatrics. Two separate researchers conducted the classification. There was complete agreement among the researchers when it comes to categorizing questions.

2.2. Data collection and analysis

Prior to presenting the questions, both ChatGPT-3.5 and ChatGPT-4 were provided with instructions outlining the examination rules, which included details such as the number of questions, the quantity of answer choices, and the count of correct answers. Furthermore, following each question, an additional query was directed to ChatGPT, inquiring, "On a scale of 1 to 5, how confident are you in this answer?" This measure was implemented to assess the level of confidence exhibited by ChatGPT in its chosen response. The confidence scale was defined as follows: 1 represented "definitely not sure," 2 "not very sure," 3 "almost sure," 4 "very sure," and 5 corresponded to "definitely sure." Each question was input into ChatGPT, and all interactions with the chat interfaces were recorded. To maintain consistency with the content of the examination questions, the conversations with the chat interfaces were conducted in the Polish language. Communication between the researchers and the two chat interfaces occurred simultaneously on separate computers, with messages sent to the chat interfaces containing identical content.

2.3. Statistical analysis

The analysis was conducted using Statistica 13.0 software (StatSoft, Krakow, Poland). The difference in ChatGPT's performance was analyzed using the exact McNemar test. Confidence intervals for proportions were computed using the Wilson score interval. The Cohen's Kappa coefficient was calculated to assess the agreement between LLMs. Categorical data is presented as counts and/or percentages, while proportions are shown as fractions and/or percentages with 95% confidence intervals. P-values of less than 0.05 were considered as significant.

3. Results

Table 1 illustrates the proportion of correct responses among the language models. GPT-4 achieved a correct response rate of 63.3% (62 out of 98 questions), whereas GPT-3.5 achieved a correct response rate of 37.8% (37 out of 98 questions). Unlike GPT-3.5, GPT-4 reached the passing rate for the NSE. There was a statistically significant ($p<0.001$) difference between dependent proportions (25,51% (95% CI: 11,97% to 39,05%)). The probability (p1) that ChatGPT-4 will provide a correct response when GPT-3.5 gives an incorrect one is 55.74% (CI_{p1}= 43.32% -67.5%). The probability (p2) that GPT-4 will provide a correct response when GPT-3.5 also gives a correct response is 75.68% (CI_{p2}= 59.90% - 86.64%). Cohen's Kappa is equal to 0.176 (95% CI: 0.009 to 0.343), indicating weak agreement between GPT-4 and GPT-3.5.

Tables 2-6 display LLM performance categorized by question category. Statistically significant ($p=0.0433$) difference between dependent proportions (18.6% (95% CI: -2.1% to 39.31%)) was observed solely in the Clinical & Case question category meaning that fraction of correct answers is different between two GPTs. The highest agreement between models is observed for questions from the Clinical & Case group (Cohen's $k = 0.458$; 95% CI: 0.214 to 0.702).

Tables 7 and 8 display the performance of LLMs based on their confidence level. The probability (p3) of providing a correct answer when the confidence level is "definitely sure" is 43.9% (95% CI: 29.9% - 59%) for GPT-3.5 and 88.9% (95% CI: 56.5% - 98%) for GPT-4. The probability (p4) of giving the correct answer when the confidence level is other than "definitely sure" is 33.3% (95% CI: 22.5% - 46.3%) for GPT-3.5 and 60.7% (95% CI: 50.3% - 70.2%) for GPT-4.

Table 1. Overall proportion of correct and incorrect answers (McNemar's test: chi-squared = 13.4; df=1; p <0.001; OR (95% CI) = 3.78 (1.78-8.96))

LLM	GPT-4			
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	28 (28.6%)	9 (9.2%)	37 (37.8%)
	No	34 (34.7%)	27 (27.6%)	61 (62.2%)
	Column total	62 (63.3%)	36 (36.7%)	N = 98

P_A= (a+b)/N= 37/98= 0.3776 (0.2879, 0.4764);
P_B= (a+c)/N= 62/98= 0.6327 (0.5339, 0.7214);
$P_B - P_A$= 0.2551 (0.1197, 0.3905);
p_1= 34/61= 0.5574 (0.4332, 0.6750);
p_2= 28/37= 0.7568 (0.5990, 0.8664);
Cohen's k: 0.1758 (0.009, 0.343).

Table 2. Distribution of correct and incorrect answers for Physiology & Diagnostics question category (McNemar's test: chi-squared = 2.5; df=1; p=0.1138; OR (95% CI) = 4 (0.798-38.666))

LLM	GPT-4			
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	6 (28.6%)	2 (9.5%)	8 (38.1%)
	No	8 (38.1%)	5 (23.8%)	13 (61.9%)
	Column total	14 (66.7%)	7 (33.3%)	N = 21

P_A= (a+b)/N= 8/21= 0.381 (0.2075, 0.5912);
P_B= (a+c)/N= 14/21= 0.6667 (0.4537, 0.8281);
P_B - P_A= 0.2857 (-0.0038, 0.5752);
p_1= 8/13= 0.6154 (0.3552, 0.8229);
p_2= 6/8= 0.75 (0.4093, 0.9285);
Cohen's k: 0.118 (-0.235, 0.471).

Table 3. Distribution of correct and incorrect answers for Clinical & Case Questions question category (McNemar's test: chi-squared = 4.083; df=1; p=0.0433; OR (95% CI)= 5 (1.066-46.933))

LLM	GPT-4			
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	17 (39.5%)	2 (4.7%)	19 (44.2%)
	No	10 (23.3%)	14 (32.6%)	24 (55.8%)
	Column total	27 (62.8%)	16 (37.2%)	N = 43

P_A= (a+b)/N= 19/43= 0.4419 (0.3043, 0.5889);
P_B= (a+c)/N= 27/43= 0.6279 (0.4786, 0.7562);
$P_B - P_A$= 0.186 (-0.0211, 0.3931);
p_1= 10/24= 0.4167 (0.2447, 0.6117);
p_2= 17/19= 0.8947 (0.6860, 0.9706);
Cohen's k: 0.458 (0.214, 0.702).

Table 4. Distribution of correct and incorrect answers for Treatment & Pharmacology question category (McNemar's test: chi-squared = 1.5; df=1; p=0.2207; OR (95% CI) = 5 (0.559-236.488))

LLM	GPT-4			
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	2 (20%)	1 (10%)	3 (30%)
	No	5 (50%)	2 (20%)	7 (70%)
	Column total	7 (70%)	3 (30%)	N = 10

P_A= (a+b)/N= 3/10 = 0.3 (0.1078, 0.6032);
P_B= (a+c)/N= 7/10 = 0.7 (0.3968, 0.8922);
$P_B - P_A$= 0.4 (-0.0017; 0.8017);
p_1= 5/7= 0.7143 (0.3589, 0.9178);
p_2= 2/3= 0.6667 (0.2077, 0.9385);
Cohen's k: -0.034 (-0.492, 0.423).

Table 5. Distribution of correct and incorrect answers for Surgery question category (McNemar's test: chi-squared = 1.125; df=1; p=0.2888; OR (95% CI) = 3 (0.536-30.393))

LLM	GPT-4		
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	1 (7.7%)	2 (15.4%)	3 (23.1%)
	No	6 (46.2%)	4 (30.8%)	10 (76.9%)
	Column total	7 (53.9%)	6 (46.2%)	N = 13

P_A= (a+b)/N= 3/13= 0.2308 (0.0818, 0.5026);
P_B= (a+c)/N= 7/13= 0.5385 (0.2914, 0.7679);
$P_B - P_A$= 0.3077 (-0.0471, 0.6625);
p_1= 6/10= 0.6 (0.3127, 0.8318);
p_2= 1/3= 0.333 (0.0615, 0.7923);
Cohen's k: -0.182 (-0.626, 0.262).

Table 6. Distribution of correct and incorrect answers for Pediatrics question category (McNemar's test: chi-squared = 0.571; df=1; p=0.4497; OR (95% CI) = 2.5 (0.409-26.253))

LLM	GPT-4		
GPT-3.5	Correct answer	Yes	No	Row total
	Yes	2 (18.2%)	2 (18.2%)	4 (36.4%)
	No	5 (45.5%)	2 (18.2%)	7 (63.6%)
	Column total	7 (63.6%)	4 (36.4%)	N = 11

P_A= (a+b)/N= 4/11= 0.3636 (0.1517, 0.6462);
P_B= (a+c)/N= 7/11= 0.6364 (0.3538, 0.8483);
$P_B - P_A$= 0.2727 (-0.1293, 0.6748);
p_1= 5/7= 0.7143 (0.3589, 0.9178);
p_2= 2/4= 0.5 (0.15, 0.85);
Cohen's k: -0.185 (-0.708, 0.338)

Table 7. Distribution of correct/false answers allocated for level of confidence for GPT3.5

Level of confidence	GPT-3.5	
	Correct	Incorrect
Definitely sure	18	23
Very sure	19	38
Almost sure	-	-
Not very sure	-	-
Definitely not sure	-	-

p3=(18/41)= 0.4390 (0.2989, 0.5896)
p4=(19/57)= 0.3333 (0.2249, 0.4628)

Table 8. Distribution of correct/false answers allocated for level of confidence for GPT4

Level of confidence	GPT-4	
	Correct	Incorrect
Definitely sure	8	1
Very sure	40	19
Almost sure	14	16
Not very sure	-	-
Definitely not sure	-	-

p3=(8/9)= 0.8889 (0.5650, 0.9801)
p4=(54/89)= 0.6067 (0.5029, 0.7018)

Table 9 displays the proportion of certainty levels assigned by LLMs. The number of observed agreements was 43, accounting for 43.88% of the observations. Cohen's Kappa coefficient was calculated to be 0.082 (95% CI: -0.018 to 0.182) meaning very week agreement between models.

4. Discussion

When comparing the accuracy of the answers using both GPT-3.5 and GPT-4.0 by question category, we discovered that the fields of surgery and pediatrics were the most lacking in terms of answer correctness. The lower accuracy in answering questions in the field of surgery and pediatric ophthalmology could be attributed to several factors, which is important to consider when analyzing the performance of language models like GPT-3.5 and GPT-4.0 in specific domains.

If the training data used for fine-tuning the models did not adequately cover pediatric ophthalmology and surgery, the models might not have gained a robust understanding of the intricacies and nuances specific to these fields, which resulted in a lesser understanding of the question asked and ultimately – an incorrect answer. Pediatric ophthalmology and surgery are highly specialized fields with complex and detailed knowledge requirements, which, if the training data did not sufficiently capture the diversity of cases and scenarios within, might be the cause of the model's struggles to provide accurate answers.

Table 9. Distribution of certainty levels between LLMs

LLM	GPT-4						
GPT-3.5	Level of confidence	Definitely sure	Very sure	Almost sure	Not very sure	Definitely not sure	Total
	Definitely sure	2 (2.04 %)	18 (18.37%)	21 (21.43%)	-	-	41 (41.84%)
	Very sure	7 (7.14%)	41 (41.84%)	9 (9.18%)	-	-	57 (58.16%)
	Almost sure	-	-	-	-	-	-
	Not very sure	-	-	-	-	-	-
	Definitely not sure	-	-	-	-	-	-
	Total	9 (9.18%)	59 (60.20%)	30 (30.61%)	-	-	N = 98

Another aspect worth noting is the question of complex decision-making processes within the NSE questions in the field of surgery. If the models lack exposure to a diverse set of surgical cases and the associated decision-making considerations, their accuracy in these scenarios may be compromised.

The matter of considerable significance also lies in the elevated specialization of medical terminology, particularly within the realm of surgery. This domain is characterized by a high degree of nuance, frequently encompassing intricate and technical language. It is plausible that the models were not specifically fine-tuned to adequately account for such nuanced linguistic intricacies in the context of surgery. This facet, although deliberated within the scope of surgical inquiries, may be extrapolated to encompass all queries embedded in the NSE, consequently yielding test outcomes lower than anticipated.

In Moshirfar et al.'s study, GPT-4 outperformed both GPT-3.5 and humans on ophthalmology questions sourced from a professional-level question bank. GPT-4 performed significantly better than GPT-3.5 (73.2% vs. 55.46%). It is important to note that in this study the subcategory "lens and cataract" presented a unique challenge for the GPT-4 and GPT-3.5 models [4].

In Cai et al.'s study, Bing Chat (Microsoft), ChatGPT 3.5, and ChatGPT 4.0 (OpenAI) were evaluated using 250 questions from the Basic Science and Clinical Science Self-Assessment Program. While ChatGPT's training data is up to 2021, Bing Chat utilizes a more up-to-date internet search to produce its responses. This was measured against the performance of human participants. Human participants achieved an accuracy rate of 72.2%. ChatGPT-3.5 had the lowest accuracy at 58.8%, while both ChatGPT-4.0, with 71.6%, and Bing Chat, with 71.2%, showed similar performance levels [5].

Integrating AI language models like GPT into healthcare is a complex task with significant challenges that may take time to overcome. Ensuring data privacy is crucial, as medical information is highly sensitive and must adhere to strict regulations. Additionally, the quality and specificity of training data are vital, yet most AI models are trained on datasets not tailored for specialized medical use, which often excludes critical medical knowledge due to accessibility restrictions.

The rapid evolution of medical knowledge presents a unique challenge for AI. Research, treatments, and protocols are continually updated, but AI models like GPT are typically trained on datasets that might not include the latest medical information. This lag can lead to outdated or incorrect recommendations from the AI, which can have serious implications in a healthcare setting. Moreover, a lot of essential clinical data is sequestered behind paywalls or restricted to certain legal and ethical guidelines, limiting the availability of the most current information for AI training.

Ethical concerns are paramount, especially regarding patient consent for using their data, the role of AI in critical decision-making, and ensuring transparency in AI-driven processes. Despite the potential benefits, the application of AI like GPT in healthcare involves navigating technical, regulatory, ethical, and societal hurdles. The journey towards integrating AI effectively in healthcare necessitates ongoing research and careful consideration of its broad impacts.

5. Limitations

Comparison of ChatGPT 3.5 and GPT-4.0 in solving a Neurosurgery Specialization Examination reveals several limitations. The reproducibility of results could be affected by the inherent variability in generating responses, even when inputs are identical. Furthermore, NSE consist solely of closed, single-choice questions, which are not necessarily indicative of the complexities encountered in real-world clinical scenarios. The evaluation criteria and metrics currently used fail to capture the intricate nuances of medical reasoning and decision-making inherent in clinical practice, indicating that these tools cannot yet be autonomously integrated into medical practice. Additionally, potential biases in the datasets used for training these models may affect their performance and generalizability across diverse clinical scenarios. Physicians need to remain aware of ongoing AI research, as its application has the potential to elevate future standards of medical care, despite currently existing limitations and ethical concerns related to the use of ChatGPT in clinical practice.

ORCID

Marcin Ciekalski https://orcid.org/0000-0003-1392-2007
Maciej Laskowski https://orcid.org/0009-0005-5809-0875
Agnieszka Koperczak https://orcid.org/0009-0003-6205-3669
Maria Śmierciak https://orcid.org/0009-0001-7563-3940
Sebastian Sirek https://orcid.org/0000-0002-3138-3011

Conflict of interest

There is no conflict of interest.

Authors' contribution

Marcin Ciekalski: Research concept and design, Carrying out the experiments, Acquisition of data, Data analysis and interpretation, Writing - Original Draft Preparation, Writing - Review and Editing. Maciej Laskowski: Research concept and design, Carrying out the experiments, Acquisition of data, Writing - Original Draft Preparation, Literature review. Agnieszka Koperczak: Carrying out the experiments, Acquisition of data, Writing - Original Draft Preparation. Maria Śmierciak: Carrying out the experiments, Acquisition of data, Writing - Original Draft Preparation. Sebastian Sirek: Research concept and design, Supervising the project, Writing - Review and Editing. All authors have approved the submission and publication of the manuscript.

References

[1] Deloitte Malta [Internet]. [cited 2023 Oct 21]. The Age of Artificial Intelligence: A brief history... | Deloitte Malta | RPA & AI. Available from: https://www2.deloitte.com/mt/en/pages/rpa-and-ai/articles/mt-age-of-ai-1-a-brief-history.html

[2] Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane PA, Wong TY. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019 Feb;103(2):167–75.

[3] Li Z, Wang L, Wu X, Jiang J, Qiang W, Xie H, Zhou H, Wu S, Shao Y, Chen W. Artificial intelligence in ophthalmology: The path to the real-world clinic. Cell Rep Med. 2023 Jul 18;4(7):101095.

[4] Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus. 2023 Jun;15(6):e40822.

[5] Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, Alabiad C. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol. 2023 Oct;254:141–9.