ASTHMA
ALLERGY
IMMUNOLOGY
ASTIM
ALLERJİ
İMMÜNOLOJİ

doi: 10.21911/aai.2024.606
Asthma Allergy Immunol 2024;22:225-236

REVIEW

Received: 05.07.2024 • Accepted: 09.08.2024
Online Published: 09.12.2024

Performance of Large Language Models in Medical Exams: A Review Focusing on Allergy, Immunology, and Related Fields

Betul DUMANOGLU[1], Ozge CAN BOSTAN[2], Onur Can YILDIRIM[3], Pamir CERCI[4]

[1] Department of Allergy and Immunology, Ağrı Training and Research Hospital, Ağrı, Türkiye
[2] Department of Allergy and Immunology, Çanakkale Mehmet Akif Ersoy State Hospital, Çanakkale, Türkiye
[3] Department of Internal Medicine, Division of Immunology and Allergy, Ege University School of Medicine, İzmir, Türkiye
[4] Division of Immunology and Allergy, Eskisehir City Hospital, Department of Internal Medicine, Eskişehir, Türkiye

Corresponding Author: Betul Dumanoglu ✉ bdumanoglu.bd@gmail.com

ABSTRACT

The emergence of large language models (LLMs) presents significant opportunities in healthcare and medical education. This study evaluates the performance of LLMs in medical examinations, with a specific focus on allergy, immunology, and related specialties. LLMs have developed the ability to comprehend, interpret, and process language in a manner akin to humans. This advancement raises concerns about their potential role in disciplines like medicine, which require advanced cognitive skills and a deep, specialized knowledge base. Following PRISMA guidelines, our review investigates the performance of LLMs in medical tests, highlighting both their strengths and limitations. We found that LLMs demonstrate higher accuracy on English-language assessments but exhibit significant variation in performance across different medical disciplines. This underscores the need for discipline-specific training and raises ethical considerations regarding challenges in clinical reasoning and visual interpretation. Future research should address linguistic biases, develop specialized protocols, and enhance the capacity of LLMs in immunology and allergy. This study emphasizes the potential of LLMs to transform medical education and advocates for their careful integration to ensure adequate support for healthcare professionals in managing complex allergic and immunological conditions.

Keywords: Large language models, LLMs, medical exam, allergy, immunology

INTRODUCTION

Large language models (LLMs) represent a significant turning point in medical education and practice. Generative artificial intelligence and its applications in many medical sectors, including allergies and immunology, have attracted increased attention since the release of ChatGPT (Chat Generative Pre-Trained Transformer) in November 2022. This review examines the efficacy of LLMs by assessing the capabilities and obstacles of LLM models in examinations within disciplines such as allergy, immunology, and related medical professions.

Recent Advancements in LLMs

Modern LLMs, such as GPT-4 (OpenAI), Claude (Anthropic), LLaMA (Meta), and PaLM 2/Gemini Pro (Google), have shown impressive skill in processing and generating text that closely mimics human language (1). Through the use of "in-text learning," these models demonstrate an impressive ability to grasp and generalize user inputs with minimal adjustments. Compared to earlier iterations, they are far more adept in confronting natural language processing chores, including translating, question-answering, and summarizing (2). There is tremendous interest in

ORCID Betul Dumanoglu / 0000-0002-4320-2425, Ozge Can Bostan / 0000-0002-4528-5404, Onur Can Yildirim / 0009-0006-2998-4877, Pamir Cerci / 0000-0002-0844-6352

how LLMs might support medical education and practice, particularly in disciplines like allergy and immunology requiring sophisticated knowledge.

LLMs in Medical Examinations

Recent studies in medical licensing tests—including the United States Medical Licencing Examination (USMLE) and board exams in specialties, including oncology and gastroenterology—show that LLMs can perform around or above the passing level (3-5). The increasing integration of artificial intelligence into medical education and assessments prompts us to reconsider its impact, especially in specialized areas like allergy and immunology. These fields demand a thorough grasp of immune system processes and their real-world clinical applications, making the role of AI both crucial and challenging.

The Gap in Current Literature

Although numerous reviews assess the applicability of LLMs in various medical applications (6,7), several studies methodically evaluate the performance of these models in medical tests, particularly in the context of allergy and immunology. Typically, contemporary research is focused on specific objectives, such as the application of ChatGPT in dentistry or ethical considerations in healthcare (8). Still, a more comprehensive study that encompasses the efficacy of multiple models in various medical disciplines has been neglected. Additionally, most of the current material comprises opinion articles rather than empirical studies highlighting the significance of a comprehensive assessment.

Objectives and Significance

This review aims to:

1. Examine the performance of LLMs in medical exams across various fields, with a specific focus on allergy and clinical immunology, pulmonology, internal medicine, pediatrics, dermatology, and otolaryngology.

2. Evaluate the accuracy and reliability of LLMs in these exams.

3. Assess the potential of LLMs as tools for medical education and assessment in allergy and immunology.

4. Identify strengths, limitations, and future research directions.

5. Consider implications and address reliability, bias, and privacy concerns in allergy and immunology practice.

To our knowledge, this is the first review to specifically evaluate LLM performance in medical exams related to allergy and immunology. Considering the complex nature of allergic and immunological disorders, understanding LLMs' performance in these areas is critical for their potential incorporation into specialized medical education and clinical decision support systems.

As LLMs advance and exhibit remarkable proficiency in evaluating medical knowledge, it is crucial to assess their performance, especially in specialized domains, as highlighted in this article. Understanding the present capabilities of LLMs enables us to more effectively integrate AI technologies into medical education and clinical practice. This integration will facilitate the training and support of healthcare professionals in effectively managing complex allergic and immunological conditions. This study aims to rigorously evaluate the performance of LLMs in medical examinations, providing valuable insights for educators, researchers, and clinicians in allergy and immunology.

MATERIALS and METHOD

We carefully followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines when planning and creating this review, instilling confidence in the quality of the methodology.

Search Strategy

This literature search was comprehensive and detailed across multiple databases (PubMed/MEDLINE, Embase, Web of Science) on literature from May 2023 to june 2024. The broader search terms/ keywords were chosen: "large language models," "LLM," "artificial intelligence," "machine learning," "medical exam," "medical education," "allergy," "Immunology," and "clinical assessment." These words were used in different combinations. Also, we used MeSH terms and free text when applicable. The complete search strategy for PubMed/MEDLINE is provided in Figure 1.

Inclusion and Exclusion Criteria

Studies were included based on the following criteria: a) Publications assessing large language models' (LLMs) performance on medical knowledge and skill exams, in-

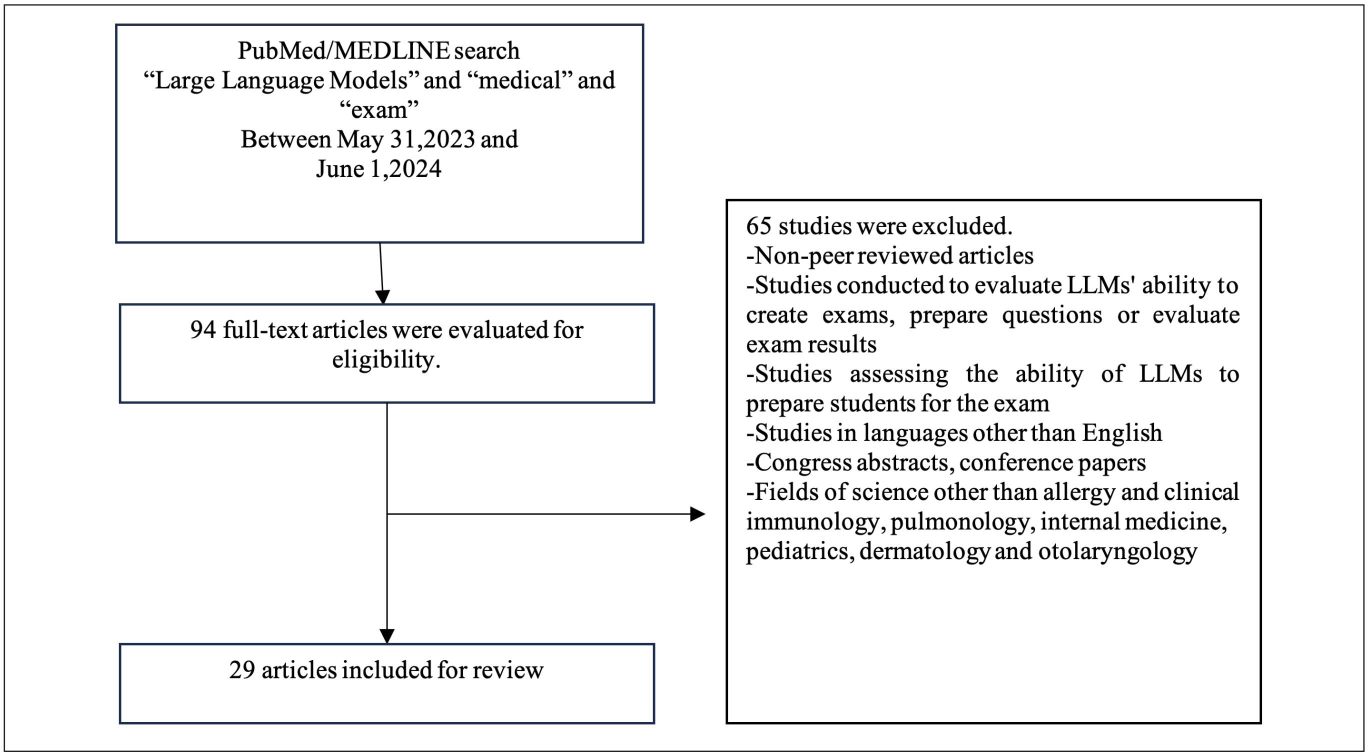
PubMed/MEDLINE search
“Large Language Models” and “medical” and “exam”
Between May 31,2023 and June 1,2024
94 full-text articles were evaluated for eligibility.
29 articles included for review
65 studies were excluded.
-Non-peer reviewed articles
-Studies conducted to evaluate LLMs' ability to create exams, prepare questions or evaluate exam results
-Studies assessing the ability of LLMs to prepare students for the exam
-Studies in languages other than English
-Congress abstracts, conference papers
-Fields of science other than allergy and clinical immunology, pulmonology, internal medicine, pediatrics, dermatology and otolaryngology

Figure 1. Searching Procedure

cluding medical specialty or license exams b) Studies comparing LLM performance on medical exams with models such as ChatGPT (OpenAI), Bard (Google), New Bing (Microsoft), Claude instant (Anthropic), Claude+ (Anthropic) and GPT-4 (OpenAI) c) Observational studies, experimental studies, and comparative analyses d) Full-text articles available in English.

We excluded: a) Abstracts not published as full-text articles b) Review articles and conference papers c) Studies focusing on LLMs’ capabilities to create exams, generate questions, or prepare students for exams d) Studies testing non-medical skills or evaluating LLMs’ exam results without comparison to human performance e) Studies not relevant to medical education or clinical assessment.

Study Selection and Data Extraction

Two highly experienced and independent reviewers meticulously screened the titles and abstracts of identified studies. Full-text articles of potentially eligible studies were then assessed. Any discrepancies were resolved thoroughly with a third reviewer.

Data extraction was performed using a standardized form. The following information was extracted: author(s), publication year, study design, LLM type, medical specialty focus (with particular attention to allergy and immunology), exam type, sample size, performance metrics, and key findings.

Quality Assessment

The quality of included studies was assessed using the Joanna Briggs Institute (JBI) Critical Appraisal tools, appropriate for each study design. For studies specific to allergy and immunology, additional consideration was given to the relevance and comprehensiveness of the exam content in representing the specialty.

Data Synthesis

Due to the anticipated heterogeneity in study designs and outcome measures, a narrative synthesis approach was adopted. We categorized findings based on the LLM type, medical specialty (focusing on allergy and immunology), and exam characteristics. Where possible, we conducted a comparative analysis of LLM performance across different specialties and human performance.

Ethical Considerations

Ethical approval was not required as this study did not involve human subjects and used only published data. However, we considered the ethical implications of LLM use in medical education throughout our analysis.

RESULTS

A total of 29 studies were included in our review. All the studies used GPT-3.5 and/or GPT-4. Google Bard, Claude, New Bing, Claude Instant, and Claude were used in only 2 studies (9,10). GPT-4 received passing grades in all the exams it was used in (9,11-16). The application of GPT-4 achieved higher scores compared to GPT-3.5 (9,11,13,15). The results are shown in Table I.

DISCUSSION

Our systematic review evaluates how large language models (LLMs) perform in medical exams, especially in fields like allergy, immunology, and related areas. We found several key points that need more attention and discussion. This analysis clarified the strengths and limitations of LLMs, particularly in the context of medical education and assessments in allergology and clinical immunology.

Table I: Literature review of exams solved with large language models

Author, Year, Reference	Method	Medical Exam	Question Number	Results
Fuchs et al., 2023, (11)	-ChatGPT-3 and 4 both -Before and after priming, -Images or illustrations were excluded -SFLEDM was translated from German to English	Swiss Federal Licensing Examination in Dental Medicine (SFLEDM), European Examination in Allergy and Clinical Immunology (EEAACI)	SFLEDM:32 EEAACI:28	-63.3% and 79.3% average accuracy for SFLEDM and EEAACI exams, respectively. -ChatGPT-4 outperformed ChatGPT-3 significantly in both assessments. -Priming improved ChatGPT-3's performance in SFLEDM and EEAACI assessments. -ChatGPT-4 showed a priming effect in SFLEDM assessment (p=0.038).
Huang et al., 2024, (12)	-ChatGPT-4 -For image-based questions, first upload the image and then enter the question. -20 questions in one conversation. -Chinese and English	Senior Professional and Technical Examinations for Medical Doctors (SPTEMD) Stage 1 in Taiwan	Feb 2022 July 2022 Feb 2023 3*200	-87.8% average accuracy -Biochemistry had the highest score at 93.8% -Anatomy, parasitology, embryology lowest -Variability in the accuracy of exam results for embryology and parasitology was observed across different tests
Le and Davis, 2024, (13)	-ChatGPT-3.5 and ChatGPT-4.0 -Comparing with passing scores (70%)	PREP® (The Pediatrics Review and Education Program) Self-Assessment, American Academy of Pediatrics (AAP) 2021 and 2022	2021 PREP®:243 2022 PREP®:247	-For chatGPT-3.5: 143 of 243 (58.85%) and 137 of 247 (55.46%) average accuracy for 2021 and 2022 PREP, respectively. -ChatGPT-4.0 correctly answered 193 of 243 (79.84%) and 208 247 (84.21%) questions. -ChatGPT-3.5's performance declined with involving media, whereas ChatGPT-4.0 showed minimal impact -ChatGPT-3.5 struggled with prompts containing tables (51.51% correct), especially those with multiple tables or more than two columns (0% correct). -In contrast, ChatGPT-4.0 handled basic tables without significant issues.

Table I continue

Watari et al., 2024,(14)	-GPT-4 -Audio, video, or images were excluded -The results were analyzed comparatively across four categories, including seven specific areas within the category of specific diseases and questions' difficulties -In the Japanese language	General Medicine In-Training Examination (GM-ITE) 2020, 2021 and 2022	137	-GPT-4 achieved a notably higher overall score than Japanese residents (GPT-4: 70.1%; residents: 55.8%; P<.001). -GPT-4 outperformed in specific disease knowledge, obstetrics, and internal medicine areas. -GPT-4's performance was lower in medical interviewing, general practice, and psychiatry. -Residents performed better on easier questions (P<.001) -Residents performed better on questions where not only knowledge but also understanding how to apply this knowledge in specific contexts was required
Meyer et al., 2024,(15)	-GPT-3.5 and GPT-4 -Visual contents were excluded (105 questions) -All questions and answers were presented to Chat GPT in the same session -In German -Passing score: 60%	German medical licensing exams, October 2021, April 2022, and October 2022	835	-Overall, GPT-4 answered 27% more questions correctly than GPT-3.5 when all exams were considered (%85 vs. %58). -GPT-4 passed all the exams, while GPT-3.5 passed only one exam -Internal medicine or surgery were more likely to be answered correctly by GPT-4. -Psychiatry was more likely to be answered correctly by GPT-3.5
Ozeri et al., 2024, (21)	-Evaluation of GPT-3.5's scores in the Hebrew National Internal Medicine Exam -In the Hebrew language	Israel National Internal Medicine Exam (Shlav Aleph) 2023	133	-GPT-3.5 correctly answered only 50 questions (36.6%). -The low overall scores may be attributed to the exam being in Hebrew.
Bielówka et al., 2024, (26)	-GPT-3.5 -Graphical content was excluded -Allergology-specific -Passing score: 60% -In Polish	Polish National Specialist Examination (PES) in Allergology Spring 2023	118	-GPT-3.5 failed the exam with a score of 52.5% (62/118). -It scored below 50% in the category requiring comprehension and critical thinking (27/60). -It performed statistically significantly better on difficult questions.
Behrmann et al., 2023,(18)	-GPT-3.5 -Only the text part of the visual questions was put in -Each question was answered one by one in different sessions	Amboss question bank (dermatology-specific questions)	492	-Only 41% of the questions were answered correctly (204/492), significantly lower than the USMLE (60%). -It answered questions 8% more accurately without images, but this was not statistically significant. -Easier questions were statistically answered more accurately.
Abbas et al., 2024, (9)	-GPT-4, GPT-3.5, Google Bard, and Claude	National Board of Medical Examiners (NBME) Exam	163	-LLM accuracy scores: -100% for GPT-4 -82.2% for GPT-3.5 -75.5% for Bard -84.7% for Claude -GPT-4's performance was statistically superior to the others -Other LLMs performances were similar

Table I continue

Nakao et al., 2024, (28)	-Assess the accuracy of GPT-4(Vision) by evaluating questions with visuals and then only their text portions -Compare the results from both scenarios -2 categories of questions: clinical (98) and general (10) -In Japanese	117th Japanese National Medical Licensing Exam (400 questions)	108 (had 1 or more images as part of a question)	-GPT-4V demonstrated an accuracy rate of 68% (73/108) with images and 72% (78/108) without images -In two categories, clinical and general, the accuracy rates were 71% (70/98) with images versus 78% (76/98) without images for clinical questions and 30% (3/10) with images versus 20% (2/10) without images for general questions. -Supplementary visual information did not significantly affect GPT-4V's performance
Zong et al., 2024,(29)	-GPT-3.5 -in the Chinese Language	National Medical Licensing Examination (NMLE), National Pharmacist Licensing Examination (NPLE), and National Nurse Licensing Examination (NNLE)	NMLE: 150*4=600 NPLE: 120*4=480 NNLE: 120*2=240	-GPT failed to achieve a passing grade in all three examinations. -This outcome is attributed to GPT's greater English proficiency and familiarity with healthcare practices in English-speaking countries, whereas the study was conducted in Chinese.
Morreel et al., 2024,(10)	-ChatGPT, Bard, New Bing, Claude instant, Claude+ and GPT-4 -Questions were translated into English.	Antwerp University multiple-choice medical license exam	102	-All utilized LLMs passed the examination. -GPT-4 and Bing achieved significantly higher results. -No significant difference was found between Claude+ and Claude Instant. -Regarding question content, no significant difference was found between clinical and theoretical questions. -Although no LLMs refused to answer clinical questions altogether, Claude+ and Claude Instant declined to answer some questions and terminated the session.
Mahajan et al., 2023,(30)	-GPT-3.5 -Accuracy and sufficiency of explanations were assessed -Otolaryngology field	-BoardVitalsTM (the question bank) -118 excluded because of images	1088	-GPT-3.5 has a 53% accuracy rate for answers and a 54% for explanations. -As the questions become more challenging, both the accuracy of answers and the clarity of explanations tend to decrease
Weng et al., 2023, (22)	-GPT -3.5 -In Chinese	-Taiwan's 2022 Family Medicine Board Exam	125	-GPT-3.5 correctly answered 52 out of 125 questions (41.6%). -This performance may be attributed to the complexity of the exam and the scarcity of Chinese language databases. -The length of the questions did not impact the accuracy rates. -Specifically, the accuracy rates were 45.5% for negative-phrase questions, 33.3% for multiple-choice questions, 58.3% for mutually exclusive options, 50.0% for case scenario questions, and 43.5% for questions related to Taiwan's local policies.

Table I continue

Noda et al., 2024, (23)	-GPT-4V -First in Japanese -Secondly, in English translation -Including 46 image-based questions	-2023 Otolaryngology board certification exam	100	-For text questions, the accuracy rate increased from 24.7% to 47.3% when translated to English and added prompts (P<.001). -GPT gave more correct answers to text questions than to visual ones -Changing the language to English and providing some instructions increased the accuracy rate -Adding visuals to the visual questions increased the accuracy rate, but no significant difference was observed (from 30.4% to 41.3% (P=.02).
Shieh et al., 2024, (16)	-GPT-3.5 and GPT-4.0, -Ability to generate a differential diagnosis -Image questions excluded	-USMLE step 2 (June 2022)	109	-GPT-4 exhibited a 40% higher accuracy rate compared to GPT-3.5, achieving 87.2% versus 47.7%. -Out of the 109 questions, 63 were case-based, and GPT-4 successfully generated an accurate differential diagnosis list for 47 of these cases
Oztermeli and Oztermeli 2023,(31)	-GPT-3.5 -Basic sciences and clinical sciences -Questions containing visual elements were excluded -In Turkish	-Medical specialty exams (MSE) last 5 years 2021 Spring, Fall 2022 Spring, Fall 2023 Spring	1177	-GPT's exam success rates ranged from 54.3% to 70.9%. -Similar accuracy scores were observed for both clinical and basic science questions. -GPT provided statistically significantly more correct answers to questions with shorter stems
Wójcik et al., 2023, (32)	-GPT-4 -In Polish	-PES (Państwowy Egzamin Specjalizacyjny)	120	-GPT-4 answered 80 questions correctly.
Scaioli et al., 2023, (27)	-GPT-3.5 -Clinical case and notional question -Visual content removed	-Italian State Exam for Medical Residency (SSM)	136 questions	-Impressive accuracy rate of 90.44%, showing exceptional results in clinical case questions -Outperformed the majority of the medical doctors who took the exam
Aljindan et al., 2023, (33)	-ChatGPT-4	Saudi Medical Licensing Exam (SMLE)	220	-GPT-4 achieved an overall accuracy of 88.6%, excelling in easy and average questions -Maintain uniform performance across various fields
Rojas et al., 2023, (19)	-ChatGPT-3.5, ChatGPT-4, ChatGPT-4 With Vision (4V; -Each 180-question exam was answered by ChatGPT three times. -In Spanish	-EUNACOM (Examen Único Nacional de Conocimientos de Medicina), a major medical examination in Chile	-3*180=540 -mirroring the EUNACOM's structure and difficulty	-GPT-4 and 4V were significantly more successful compared to version 3.5, with accuracy rates of 79.32%, 78.83%, and 57.53%, respectively (P<.001). -The results of GPT-4 and 4V were similar. -While GPT-4 and 4V achieved the best results in surgery, version 3.5 was more successful in psychiatry.
Garabet et al., 2023, (34)	-ChatGPT-4	AMBOSS question bank for the USMLE STEP 1	1300	-GPT-4 accurately responded to 86% of all questions, showing consistent performance across different systems and disciplines -GPT-4 provided more accurate responses to the questions correctly answered by the highest-performing students
Lin et al., 2024, (35)	-ChatGPT-4 -Chain of thought prompt used -In the traditional Chinese Language	-Taiwan's medical licensing exam -February 2022, July 2022, February 2023 July 2033	4*80	-GPT-4's highest and lowest accuracy rates were 93.75% and 63.75%, respectively, achieved in the February 2022 and July 2023 exams. -After being trained with a type of "chain of thought" method, GPT-4 was able to provide correct responses to incorrect answers with an accuracy ranging from 0.00% to 88.89% -.ChatGPT-4's final results varied from 90% to 98%

Table I continue

Huang et al., 2023, (20)	-GPT-3.5, GPT-4 -2 image questions were excluded	-An official University of Toronto Department of Family and Community Medicine Progress Test	108	-The accuracy rates of Family Medicine residents and GPT-3.5 were similar, but GPT-4 was significantly better than both, with accuracy rates of 56.9%, 57.4%, and 82.5%, respectively -GPT-3.5 exhibited the poorest performance in elderly care -GPT-4 achieved superior scores in all 11 area
Takagi et al., 2023, (36)	-ChatGPT-4V vs examinees -In the Japanese language -Questions were in three different areas: Essential knowledge General clinical knowledge Specific diseases	117th JMLE (Japanese Medical Licensing Exam)	386	-When considering all questions together, examinees outscored GPT, but the difference was not statistically significant. (84.9 vs 78.2 p=0.003) -However, in the general clinical knowledge category, examinees' scores were significantly higher (83.1 vs 70.8, p<0.001) -GPT-4V performed on questions containing images and/or tables significantly lower than the examinees.
Gilson et al., 2023, (37)	-Chat gpt, GPT-3 and InstructGPT	-AMBOSS question bank for USMLE step 1 and 2 -National Board of Medical Examiners (NBME) questions	AMBOSS step1 and 2 (200) NBME-Free-Step1 (87) NBME-Free-Step2 (102)	-ChatGPT outperformed both InstructGPT and GPT-3 in terms of performance. -GPT-3 performed at a level comparable to random guessing -ChatGPT demonstrated better results in Step 1 and NBME exams compared to Step 2 and AMBOSS. -The highest accuracy rate achieved by ChatGPT was 64.4% on the NBME-Free-Step1 questions.
Lewandowski et al., 2023, (25)	-GPT-3.5 and GPT-4	Dermatology Specialty Certificate Exam Fall 2016 Spring 2017 Fall 2017	120*3	-GPT-4 demonstrated statistically significant better performance across all exams. -Although there was no significant difference between Polish and English versions for both GPT-3.5 and GPT-4, better results were achieved in the English versions
Alessandri Bonetti et al., 2023, (38)	ChatGPT-3	Italian Residency Admission National Exam	140	-GPT answered 87% of the questions correctly. -With this score, it would have been able to secure a spot in any medical specialty it desired -The explanations provided for the correct answers were satisfactory.
Tanaka et al., 2024, (24)	-GPT3.5 and GPT-4 -Images were excluded -The accuracy rates were examined: 1) after translating to English 2) after using prompts designed to guide GPT to provide correct answers	116th and 117th National Medical Licensing Examination (NMLE) in Japan	290, 262	-For the 116th NMLE, GPT-3.5 correctly answered 53.7% of the essential questions without any interventions. -Next, the questions were translated into English and GPT-3.5 was asked to answer them again, which increased the accuracy rate to 60.4%. -After applying appropriate prompts, the accuracy rate improved to 64.6%. -Finally, when the most optimized version of the questions was presented to GPT-4, a 90.9% accuracy rate was achieved. -For the 117th NMLE, under optimal prompts and translated into English, GPT-4 achieved an accuracy rate of 82.7% for this set of questions.

General Findings

LLMs appear to exhibit superior performance in English-language medical exams compared to tests in other languages. This is likely because most of the training data is in English. These language gaps show how important it is to adapt language inputs and improve translation strategies to optimize LLM efficacy in global medical education. Also, the results vary a lot across different medical fields. This shows that these models need more development to achieve high accuracy for specific contexts. For example, they still struggle with clinical reasoning and visual interpretation, which are essential for allergy and immunology practice (17).

Performance Variability Across Medical Specialties

Studies in the literature indicate disparities in the performance of LLMs across different medical specialties. Fuchs et al. indicated that LLMs attained a 79.3% accuracy rate on the European Academy of Allergy and Clinical Immunology (EAACI) examination. This outcome is notable given that immunology and allergic diseases typically necessitate intricate diagnostic and therapeutic decision-making (11).

Nonetheless, the efficacy of LLMs in other allergy-related disciplines is comparatively lower. Behrmann and colleagues observed that ChatGPT attained merely a 41% accuracy rate on dermatological inquiries. Despite dermatology's close association with allergies, this low outcome suggests that the efficacy of LLMs may differ according on the requisite knowledge and reasoning techniques in other medical disciplines (18). In allergy and immunology, this variability raises questions about the models' ability to navigate the intricate relationships between immunological mechanisms, environmental factors, and clinical presentations that characterize allergic diseases.

Watari et al. compared exam results of ChatGPT and medical residents, finding that GPT-4 excels in handling detailed disease knowledge and challenging questions, particularly in internal medicine and obstetrics/gynecology. However, it struggles with medical interviewing and psychiatry, which demand situational understanding and human empathy. This suggests that while AI can manage extensive knowledge-based queries, it still lacks the intuitive and experiential understanding required in fields like psychiatry and potentially in the nuanced patient interactions common in allergy and immunology practice (14).

Interestingly, Rojas et al. found contrasting results, with GPT-3.5 providing the best responses in psychiatry. These varying outcomes across medical specialties are attributed to the unique terminologies of each field, the way questions are formulated, and the inherently challenging nature of medicine (19). This inconsistency highlights the need for careful evaluation and specialization of LLMs for allergy and immunology education.

Huang et al. noted that LLMs like GPT-3.5 exhibited poor performance in areas such as elderly care, possibly due to the complexity and atypical presentations of conditions in geriatric patients (20). This finding underscores the need for more sophisticated AI models capable of handling medical conditions' nuanced and often multifaceted nature, particularly in specialties like allergy and immunology, where patient presentations can be highly variable and age-dependent.

Impact of Language on LLM Performance

The performance of LLMs in medical exams is notably influenced by language differences, a factor especially critical in allergy and immunology, which rely heavily on global collaboration. The predominance of English in medical literature and internet resources contributes to higher performance in English-language exams than in other languages.

Fuchs et al. directly compared LLM performance on exams in various languages in allergy and immunology (11). Their study evaluated LLM performance on the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM) and the EEAACI. The findings revealed average accuracies of 63.3% for SFLEDM (questions translated from German to English) and 79.3% for EEAACI (originally in English). This disparity suggests that even translated questions may pose challenges for LLMs compared to those composed initially in English, potentially impacting the standardization of allergy and immunology education across linguistic boundaries.

The language effect is even more pronounced in non-English exams without translation. Ozeri et al. found that ChatGPT correctly answered only 36.6% of Hebrew National Internal Medicine Exam questions (21). Similarly, Weng et al. reported a 41.6% accuracy rate for ChatGPT on Taiwan's Family Medicine Board Exam in Chinese (22). These results underscore the significant challenges LLMs face when dealing with non-English medical content, a critical consideration for the global allergy and immunology practice.

Several studies have explored strategies to mitigate language bias. Noda et al. have demonstrated that translating Japanese Otolaryngology Board Certification exam questions into English increased GPT-4V's accuracy from 24.7% to 47.3% (23). Tanaka et al. found that translating questions from Japanese to English and optimizing prompts improved LLM performance on Japan's National Medical Licensing Examination, particularly for GPT-4 (24). These findings suggest that combining language optimization and advanced LLM versions could enhance performance in complex clinical reasoning tasks, including allergy and immunology.

Rojas et al. assessed ChatGPT's performance on the EUNACOM, a major medical exam in Chile conducted in Spanish (19). ChatGPT-4 showed an overall accuracy of 88.6%, demonstrating that while LLMs may perform better in English, they can still achieve high accuracy in other languages under certain conditions. This result offers promise for applying LLMs in non-English speaking regions, potentially facilitating more equitable access to advanced educational tools in allergy and immunology worldwide.

Clinical Reasoning and Visual Interpretation

Although large language models (LLMs) exhibit remarkable proficiency in knowledge-based inquiries, their effectiveness in tasks necessitating intricate clinical reasoning or visual data interpretation remains variable. This constraint is especially pertinent in allergy and immunology, where the interpretation of skin tests, immunological assays, and imaging studies is essential to clinical practice.

In a study by Lewandowski et al., ChatGPT-4 answered clinical image-type questions with an average accuracy of 93.0% for English and 84.2% for Polish questions in dermatology exams (25). Although these results are promising, they may not fully reflect the complexity of visual interpretation needed in allergy and immunology practice, such as the nuanced assessment of skin prick tests, patch tests, or immunohistochemistry results.

Bielówka et al. classified study questions into "memory" and "comprehension and critical thinking," discovering that ChatGPT excelled in "memory" and difficult questions but exhibited a diminished success rate in critical thinking inquiries. This indicates that although the model is proficient in retrieving information and accessing stored data, it encounters difficulties with intricate reasoning and reflective judgment—abilities essential for diagnosing and planning treatment in allergy and immunology (26).

In contrast, Behrmann et al. noted that the model exhibited superior performance on simpler questions, attributing this to the presence of visual elements under challenging questions that GPT did not process (18). This limitation is particularly relevant in allergy and immunology, where visual cues frequently play a crucial role in diagnosis.

Ethical Considerations and Implementation Challenges

Integrating large language models (LLMs) into medical education and assessment—especially within the fields of allergy and immunology—raises significant ethical considerations. It is imperative that we thoroughly examine AI's role in healthcare decision-making, address potential biases in LLM training data, and tackle data privacy issues. For instance, if LLMs are predominantly trained on data from specific populations, they may fail to adequately represent the diverse presentations of allergic and immunological diseases across different ethnic groups, potentially exacerbating health inequalities in allergy treatment. A recent study in Japan demonstrated that LLMs can generate convincing explanations for incorrect responses, highlighting the critical need for experienced physicians to review GPT outputs in the medical field (24). The expertise and adaptability of physicians are crucial for making personalized and responsive treatment decisions, particularly in complex areas like allergy and immunology, where patient cases can vary significantly and treatment outcomes may be uncertain (27).

Moreover, introducing LLMs into medical education should be cautiously approached to ensure they enhance rather than replace essential components of clinical training. This is particularly important in allergy and immunology, where hands-on experience—such as performing skin prick tests, managing anaphylaxis, or conducting oral food challenges—is indispensable and cannot be fully replicated by AI systems.

Future Directions and Conclusion

This comprehensive review highlights the significant potential of large language models (LLMs) in enhancing medical education and assessment, including within the specialized fields of allergy and immunology. However, several challenges must be addressed before their widespread implementation can be recommended:

1. Mitigating Language and Cultural Biases through improving multilingual training and developing context-aware translation systems are crucial for reducing language and cultural biases in the global practice of allergy and immunology.

2. Developing specialty-specific training and evaluation protocols, particularly for complex fields like allergy and immunology, ensures LLMs can navigate the intricacies of immunological mechanisms and their clinical manifestations.

3. Enhancing clinical reasoning and visual interpretation of LLMs is essential for accurately diagnosing and treating allergic and immunological disorders.

4. Setting clear ethical guidelines for LLM use in medical education, ensuring patient privacy, and addressing potential biases are imperative to safeguard practice in allergy and immunology.

Long-term studies should focus on how the integration of LLMs affects medical education and clinical competency in allergy and immunology. Comparing LLMs with medical students and professionals in interpreting complex allergy tests, identifying severe immunological conditions, and planning treatments for challenging cases may provide valuable insights. Additionally, enhancing LLMs' ability to understand and generate content about the molecular and cellular mechanisms underlying allergic and immunological diseases could significantly advance the field.

In conclusion, while LLMs offer promising opportunities to augment medical education and assessment in allergy, immunology, and related disciplines, their implementation should be approached cautiously, considering their current limitations and potential long-term impacts on clinical practice. As these technologies evolve, ongoing evaluation and refinement will be crucial to ensure they effectively support the development of skilled healthcare professionals capable of navigating the complex landscape of allergy and clinical immunology. The potential of LLMs to revolutionize medical education is substantial; however, it must be realized through thoughtful implementation that prioritizes accuracy, equity, and the unique needs of specialties like allergy and immunology. By addressing the challenges identified in this review, we can work toward integrating advanced AI technologies into medical education and practice, ultimately enhancing learning outcomes and improving patient care in allergy and immunology.

Acknowledgments

This is a notification from the authors that there is no funding or research grants received in the course of the study, research, or assembly of the manuscript.

Conflict of Interest

The authors report no conflict of interest.

Authorship Contributions

Concept: **Ozge Can Bostan, Pamir Cerci,** Design: **Betul Dumanoglu, Onur Can Yildirim,** Data collection or processing: **Ozge Can Bostan, Onur Can Yildirim,** Analysis or Interpretation: **Betul Dumanoglu, Pamir Cerci,** Literature search: **Betul Dumanoglu, Ozge Can Bostan,** Writing: **Ozge Can Bostan, Pamir Cerci, Betul Dumanoglu, Onur Can Yildirim,** Approval: **Onur Can Yildirim, Pamir Cerci.**

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-40.
2. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access 2024.
3. Ali S, Shahab O, Shabeeb R Al, Ladak F, Yang JO, Nadkarni G, et al. General purpose large language models match human performance on gastroenterology board exam self-assessments. medRxiv 2023;2023.09.21.23295918.
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312.
5. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023;13:1219326.
6. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. J Med Educ Curric Dev 2024;11:23821205241238641.
7. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton EW, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. medRxiv [Internet]. 2024 Apr 27 [cited 2024 Jun 30]; Available from: /pmc/articles/PMC11071576/
8. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, Paiwal K, Singh R. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. Cureus 2023;15(6):e40367.
9. Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. Cureus 2024;16(3):e55991.
10. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. PLOS Digit Health 2024;3(2):e0000349.

11. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study. Swiss Dent J 2023;134(5).

12. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit Health 2024;10:20552076241233144.

13. Le M, Davis M. ChatGPT Yields a Passing Score on a Pediatric Board Preparatory Exam but Raises Red Flags. Glob Pediatr Health 2024;11:2333794X241240327.

14. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance Comparison of ChatGPT-4 and Japanese Medical Residents in the General Medicine In-Training Examination: Comparison Study. JMIR Med Educ 2023;9:e52202.

15. Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. JMIR Med Educ 2024;10:e50965.

16. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep 2024;14(1):9330.

17. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. 2023 Jun 16 [cited 2024 Jul 1]; Available from: http://arxiv.org/abs/2306.09968

18. Behrmann J, Hong EM, Meledathu S, Leiter A, Povelaitis M, Mitre M. Chat generative pre-trained transformer's performance on dermatology-specific questions and its implications in medical education. J Med Artif Intell 2023;6.

19. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Shima Salehi S. Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. JMIR Med Educ 2024;10: e55048.

20. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study. JMIR Med Educ 2023;9:e50514.

21. Ozeri DJ, Cohen A, Bacharach N, Ukashi O, Oppenheim A. Performance of ChatGPT in Israeli Hebrew Internal Medicine National Residency Exam. Isr Med Assoc J 2024;26(2):86-8.

22. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023;86(8):762-6.

23. Noda M, Ueno T, Koshu R, Takaso Y, Shimada MD, Saito C, et al. Performance of GPT-4V in Answering the Japanese Otolaryngology Board Certification Examination Questions: Evaluation Study. JMIR Med Educ 2024;10:e57054.

24. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health 2024;3(1):e0000433.

25. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol 2024;49(7):686-91.

26. Bielówka M, Kufel J, Rojek M, Mitręga A, Kaczyńska D, Czogalik Ł, et al. Evaluating ChatGPT-3.5 in allergology: performance in the Polish Specialist Examination. Alergologia Polska - Polish Journal of Allergology 2024;11(1):42-7.

27. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita 2023;59(4):267-70.

28. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ 2024;10:e54393.

29. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24(1):1-9.

30. Mahajan AP, Shabet CL, Smith J, Rudy SF, Kupfer RA, Bohm LA. Assessment of Artificial Intelligence Performance on the Otolaryngology Residency In-Service Exam. OTO Open 2023;7(4):e98.

31. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023;102(32):e34673.

32. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: Performance of ChatGPT on a PES medical examination. Cardiol J 2023;31(3):442-50.

33. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Falah Almutairi F, et al. ChatGPT Conquers the Saudi Medical Licensing Exam: Exploring the Accuracy of Artificial Intelligence in Medical Knowledge Assessment and Implications for Modern Medical Education. Cureus 2023;15(9):e45043.

34. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. Med Sci Educ 2023;34(1):145-52.

35. Lin SY, Chan PK, Hsu WH, Kao CH. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. Digit Health 2024;10:20552076241237678.

36. Takagi S, Koda M, Watari T. The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam. JMIR Med Educ 2024;10:e54283.

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

38. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? Ann Biomed Eng 2024;52(4):745-9.