European Journal of Therapeutics

pISSN: 2564-7784
eISSN: 2564-7040

Eur J Ther. 2025;31(1):35-43.
https://doi.org/10.58600/eurjther2611

Original Research

Evaluating the Performance of Large Language Models in Anatomy Education Advancing Anatomy Learning with ChatGPT-4o

Fatma Ok[1], Burak Karip[1], Fulya Temizsoy Korkmaz[1]

[1] Department of Anatomy, University of Health Sciences, Hamidiye Faculty of Medicine, Istanbul, Türkiye

Received: 2025-01-27 **Accepted:** 2025-02-18 **Published Online:** 2025-02-28

Corresponding Author

Fatma Ok, Anatomy Specialist / Lecturer

Address: University of Health Sciences, Hamidiye Faculty of Medicine, Department of Anatomy, Istanbul, Türkiye

E-mail: fatmaozbakirr@gmail.com
fatma.ok@sbu.edu.tr

ABSTRACT

Objective: Large language models (LLMs), such as ChatGPT, Gemini, and Copilot, have garnered significant attention across various domains, including education. Their application is becoming increasingly prevalent, particularly in medical education, where rapid access to accurate and up-to-date information is imperative. This study aimed to assess the validity, accuracy, and comprehensiveness of utilizing LLMs for the preparation of lecture notes in medical school anatomy education.

Methods: The study evaluated the performance of four large language models—ChatGPT-4o, ChatGPT-4o-Mini, Gemini, and Copilot—in generating anatomy lecture notes for medical students. In the first phase, the lecture notes produced by these models using identical prompts were compared to a widely used anatomy textbook through thematic analysis to assess relevance and alignment with standard educational materials. In the second phase, the generated lecture notes were evaluated using content validity index (CVI) analysis. The threshold values for S-CVI/Ave and S-CVI/UA were set at 0.90 and 0.80, respectively, to determine the acceptability of the content.

Results: ChatGPT-4o demonstrated the highest performance, achieving a theme success rate of 94.6% and a subtheme success rate of 76.2%. ChatGPT-4o-Mini followed, with theme and subtheme success rates of 89.2% and 62.3%, respectively. Copilot achieved moderate results, with a theme success rate of 91.8% and a subtheme success rate of 54.9%, while Gemini showed the lowest performance, with a theme success rate of 86.4% and a subtheme success rate of 52.3%. In the Content Validity Index (CVI) analysis, ChatGPT-4o again outperformed the other models, exceeding the thresholds with an S-CVI/Ave value of 0.943 and an S-CVI/UA value of 0.857. ChatGPT-4o-Mini met the S-CVI/UA threshold (0.714) but fell slightly short of the S-CVI/Ave threshold (0.800). Copilot and Gemini, however, exhibited significantly lower CVI results. Copilot achieved an S-CVI/Ave value of 0.486 and an S-CVI/UA value of 0.286, while Gemini obtained the lowest scores, with an S-CVI/Ave value of 0.286 and an S-CVI/UA value of 0.143.

Conclusion: This study assessed various LLMs through two distinct analysis methods, revealing that ChatGPT-4o performed best in both thematic analysis and CVI evaluations. These results suggest that anatomy educators and medical students could benefit from adopting ChatGPT-4o as a supplementary tool for anatomy lecture notes generation. Conversely, models like ChatGPT-4o-Mini, Gemini, and Copilot require further improvements to meet the standards necessary for reliable use in medical education.

Keywords: anatomy education, large language models, content validity index, medical education, thematic analysis

INTRODUCTION

Artificial Intelligence (AI) continues to reshape multiple sectors, including the field of education. Large language models (LLMs) are a specific type of AI model designed for natural language processing tasks. LLMs are now being utilized as support tools across nearly every field in today's world. ChatGPT, introduced in November 2022, is an LLM powered by AI. Trained on extensive multilingual text datasets, it can produce human-like responses [1]. Google's Gemini, was launched in December 2023, while Microsoft's Copilot, was released in March 2023. These models have garnered mixed reactions from the scientific community, recognized for their ability to improve efficiency in academic writing. LLMs also have shown remarkable utility in the healthcare field, assisting with clinical diagnoses, enhancing decision-making processes, providing tailored healthcare solutions, advancing drug development, and analyzing vast clinical datasets [2,3]. Despite its significant potential, LLM's use in medical education remains underexplored.

The roots of LLMs go back to the 1950s, when art AI became an academic field and the Georgetown–IBM experiment showed that machines could translate languages [4]. Before diving into the key developments that led to today's advanced technology, it's helpful to define some basics. A language model is a program that processes and generates human language, ranging from simple systems based on rules to complex AI models. LLMs are a special type of language model, known for their large size, complexity, and unique abilities that smaller models lack. Built using deep learning and trained on massive datasets with billions of parameters, LLMs excel at tasks like summarizing, translating, analyzing sentiment, and generating text. Essentially, they work by predicting the next word or token in a sequence of text. LLMs, such as ChatGPT-4, Gemini, and Copilot, have significantly advanced the field of AI by showcasing an extraordinary ability to interpret and generate text with human-like precision (accessed: December 12, 2024: https://chat.openai.com, https://gemini.google.com/app, https://copilot.microsoft.com). These chatbot models are built using extensive internet datasets, enabling them to absorb a wide range of knowledge and subtle language intricacies [5,6]. Since 2023 the use of artificial intelligence (AI) technologies has rapidly become a core element of every healthcare profession and represents the start of a transformative paradigm change in education with significant potential to change the way we act and teach and ultimately to improve learning outcomes. This shift requires consideration of the appropriate integration of these technologies [7,8].

Main Points

- The study utilized thematic analysis and content validity index (CVI) evaluations to comprehensively assess the effectiveness of large language models in generating accurate and relevant anatomy lecture notes.
- ChatGPT-4o demonstrated the highest performance among the evaluated large language models, proving to be the most effective tool for generating anatomy lecture notes.
- Other models, including ChatGPT-4o-Mini, Copilot, and Gemini, showed lower accuracy and content validity, indicating the need for further development before they can be reliably used to generate anatomy lecture notes.
- The study emphasizes that while LLMs can support anatomy education, they should complement, not replace, traditional teaching methods, with expert oversight ensuring accuracy..

The application of LLMs in anatomy education has begun to attract research interest. A study by Ilgaz and Çelik (2023) investigated the effectiveness of ChatGPT and Google Bard in generating anatomy-related content, finding that while LLMs could generate quizzes and provide general information, their accuracy in article writing remained inconsistent [9]. Similarly, Arun et al. (2024) compared ChatGPT with a customized AI chatbot (Anatbuddy) designed specifically for anatomy education, concluding that domain-specific models with curated knowledge bases performed better than general-purpose LLMs in factual accuracy and relevance [10]. Studies have also shown that ChatGPT-4 outperforms undergraduate students in anatomy assessments and surpasses other AI models in answering medical multiple-choice questions, yet concerns remain regarding its accuracy and reliability [11,12]. These findings suggest that while LLMs offer valuable support in anatomy education, they require further refinement to match the precision of specialized learning tools.

Given the rapid integration of AI into educational workflows, it is essential to assess the effectiveness of LLMs in anatomy education. This comparative study focused on generating educational materials for anatomy educators, with the goal of assessing the validity and inclusivity of various LLMs

(ChatGPT-4o, ChatGPT-4o-Mini, Gemini, and Copilot). Using thematic analysis and content validity analysis, the study aimed to explore the quality, relevance, and applicability of the content generated by AI. Through this integrated approach, the research aimed to identify the strength and limitations of different LLM tools that could potentially assist in streamlining and improving educational workflows in anatomy. Moreover, this study sought to further insight into the practical viability of LLMs even in contemporary anatomy educational practices.

MATERIAL AND METHODS

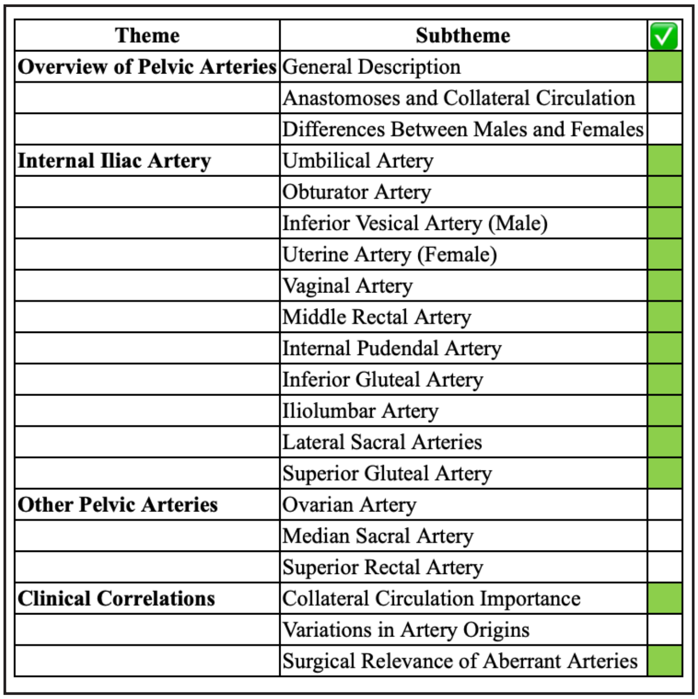
Study Design

In the current study, thematic analysis, a method used to evaluate qualitative data, and the Content Validity Index (CVI), a quantitative measurement based on expert evaluation, were employed. These complementary methods allowed for the evaluation of the anatomy lecture notes provided by LLMs using two distinct approaches.

Thematic Analysis

Thematic analysis is a qualitative research method used to identify, analyze, and interpret patterns or themes within data. Researchers systematically code the data to capture recurring ideas and organize them into themes that reflect underlying meanings or significant trends. It is widely used for its flexibility in uncovering insights across diverse datasets, including interviews, texts, and observations. Thematic analysis was employed to categorize the contents of the anatomy textbook into meaningful groups (themes and subthemes).

In the initial phase of the study, a framework for thematic analysis was established using the widely utilized anatomy textbook [13]. From the anatomical regions covered in the textbook—thorax, abdomen, pelvis and perineum, back, lower limb, upper limb, and head—one chapter was randomly selected from each of the seven sections. Each selected chapter was thoroughly reviewed from the textbook, and themes and sub-themes were developed based on this review. The themes and sub-themes were finalized through consensus between two independent researchers. In cases of disagreement, a third researcher was consulted to resolve conflicts. Subsequently, the applications ChatGPT-4o, ChatGPT-4o-mini, Gemini, and Copilot were prompted with the command, "Prepare a detailed lecture note explaining the anatomy of the ... for medical students." The generated outputs were then compared to the thematic and sub-thematic framework developed in the earlier phase (Figure 1).



Theme	Subtheme	✅
Overview of Pelvic Arteries	General Description	🟩
	Anastomoses and Collateral Circulation	
	Differences Between Males and Females	
Internal Iliac Artery	Umbilical Artery	🟩
	Obturator Artery	🟩
	Inferior Vesical Artery (Male)	🟩
	Uterine Artery (Female)	🟩
	Vaginal Artery	🟩
	Middle Rectal Artery	🟩
	Internal Pudendal Artery	🟩
	Inferior Gluteal Artery	🟩
	Iliolumbar Artery	🟩
	Lateral Sacral Arteries	🟩
	Superior Gluteal Artery	🟩
Other Pelvic Arteries	Ovarian Artery	
	Median Sacral Artery	
	Superior Rectal Artery	
Clinical Correlations	Collateral Circulation Importance	🟩
	Variations in Artery Origins	
	Surgical Relevance of Aberrant Arteries	🟩

Figure 1. Comparison of the themes and subthemes of ''pelvic arteries'' derived from the ''pelvis and perineum'' section of the anatomy textbook, with the content provided in the lecture notes generated by the Gemini application. Green areas represent themes and subthemes where the information was adequately covered.

Content Validity Analysis (CVI)

Content validity index (CVI) is a quantitative measure used to assess the degree to which items in a measurement tool, are relevant and representative of the concept or construct being evaluated. It is commonly used in the development and validation of instruments in fields like education, psychology, and healthcare.

The LLMs were tasked with generating content for selected topics from an anatomy textbook as in the initial phase of the study. Standardized prompts were provided to ensure consistency in the structure and focus of the generated materials. The content generated by the ChatGPT-4o, ChatGPT-4o-mini, Gemini, and Copilot, based on the provided prompts (''Prepare a detailed lecture note explaining the anatomy of the ... for medical students.''), was evaluated and scored by experts. Each expert reviewed the generated content for missed information, relevance, clarity, and accuracy, assigning scores on a scale of 1 to 4. The contents generated by the LLMs were scored by five experts, each holding either a Ph.D. in anatomy or a medical doctor specializing in the field of anatomy. Based on the scores provided, S-CVI/Ave and S-CVI/UA were calculated to assess the content validity of the generated anatomy lecture notes. S-CVI/Ave represents the average proportion of items rated as relevant by experts across the entire scale, while S-CVI/UA reflects the proportion of items where all experts unanimously agree on their relevance. The threshold values for S-CVI/Ave and S-CVI/UA have been established as 0.90 and 0.80, respectively, to evaluate the content validity of the lecture notes generated by LLMs [14]. All figures presented in this study were generated with the assistance of ChatGPT-4o.

RESULTS

The topics of breast, large intestine, pelvic arteries, extrinsic back muscles, hip bone, bones of the hand, and nerves of the orbit were randomly selected from the seven designated anatomical regions in the anatomy textbook: thorax, abdomen, pelvis and perineum, back, lower limb, upper limb, and head [13]. The analysis identified a total of 37 themes and 151 subthemes. ChatGPT-4o demonstrated the highest performance, achieving a theme success rate of 94.6% and a subtheme success rate of 76.2%. ChatGPT-4o-Mini achieved a theme success rate of 89.2% and a subtheme success rate of 62.3%. Copilot exhibited theme and subtheme success rates of 91.8% and 54.9%, respectively, while Gemini achieved theme and subtheme success rates of 86.4% and 52.3%, respectively (Figure 2).

The CVI analysis results revealed that among the evaluated modes, ChatGPT-4o demonstrated the highest performance with an S-CVI/Ave value of 0.943 and an S-CVI/UA value of 0.857, both exceeding the acceptable thresholds of 0.90 and 0.80, respectively. ChatGPT-4o-Mini followed with an S-CVI/Ave value of 0.800 and an S-CVI/UA value of 0.714, meeting the S-CVI/UA threshold but falling slightly below the S-CVI/Ave threshold. Copilot achieved an S-CVI/Ave value of 0.486 and an S-CVI/UA value of 0.286, both significantly below the acceptable thresholds. Similarly, Gemini exhibited the lowest performance, with an S-CVI/Ave value of 0.286 and an S-CVI/UA value of 0.143, failing to meet either threshold (Figure 3).

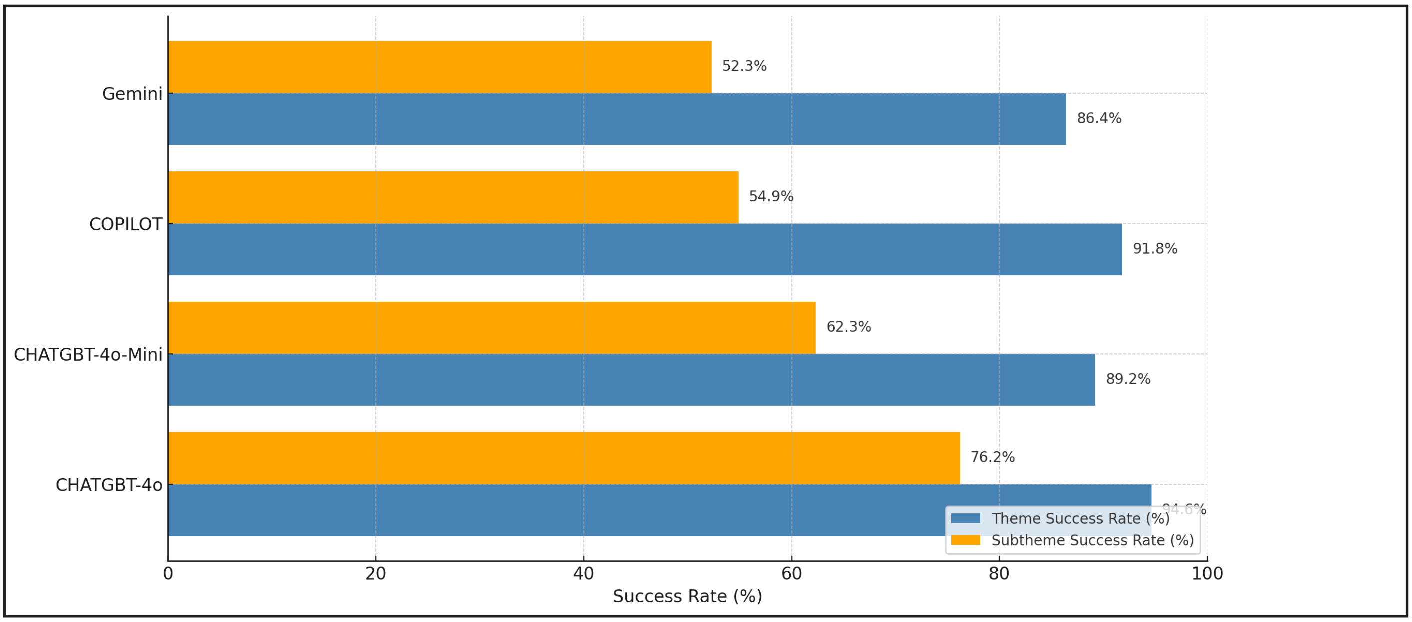


Figure 2. Comparison of theme and subtheme success rates across different models. The horizontal bars represent the performance of each model in terms of theme success rate (blue bars) and subtheme success rate (orange bars).

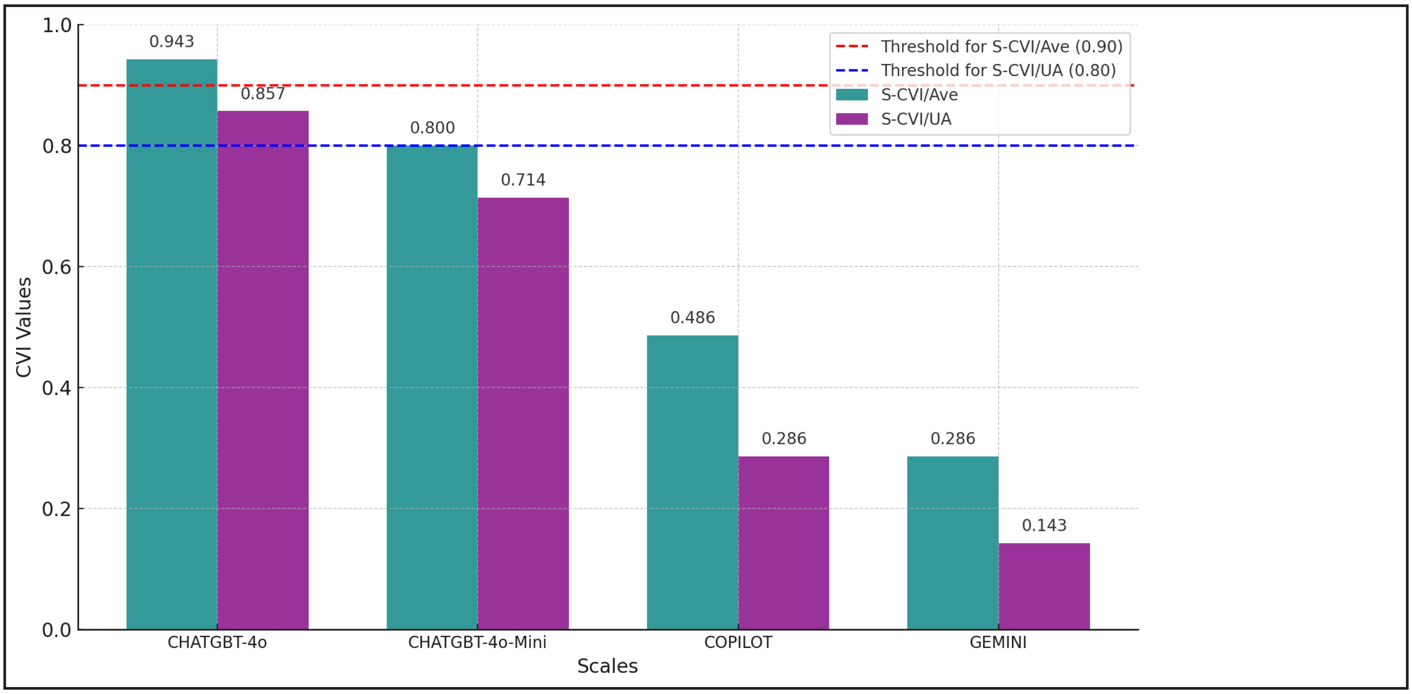


Figure 3. Comparison of S-CVI/Ave and S-CVI/UA values across different models, with thresholds (0.90 for S-CVI/Ave and 0.80 for S-CVI/UA) indicated by dashed lines.

DISCUSSION

Large Language Models (LLMs) continue to gain popularity across various domains, including medical education. These models have the potential to serve as useful tools for both students and educators by generating easily accessible and highly specific information. In particular, LLMs can be leveraged to support educators teaching in medical schools by enhancing their ability to prepare and deliver effective lessons. This study investigated the integration of LLM-based applications into anatomy education, focusing on their usability, benefits, and challenges for educators.

In this study, the initial thematic analysis revealed that ChatGPT-4o demonstrated the highest performance, achieving a theme success rate of 94.6% and a subtheme success rate of 76.2%. As the most widely used and popular LLM, ChatGPT-4o showed strong performance in covering overarching topics comprehensively. However, its relatively lower success rate in subthemes suggests that while it effectively addresses the broad outlines of topics, it provides less detailed information. Despite this limitation, ChatGPT-4o appears to be the suitable model currently available for preparing anatomy lecture notes. Comparative analysis of other models showed the following results: ChatGPT-4o-Mini achieved a theme success rate of 89.2% and a subtheme success rate of 62.3%. Copilot exhibited theme and subtheme success rates of 91.8% and 54.9%, respectively, while Gemini achieved theme and subtheme success rates of 86.4% and 52.3%, respectively. A notable trend across all models was the high performance in capturing thematic content, whereas the lower subtheme success rates reflected a relative deficiency in generating detailed, nuanced content. This suggests a clear distinction in the applicability of these models. While they excel at creating slide headings and summarizing overarching themes, the limited depth of content highlights the need for educators to supplement these outputs with additional details and context. Therefore, these models may serve as effective tools for generating outlines or slide titles, but educators should ensure that the content is enriched with supplemental information to provide comprehensive learning materials.

ChatGPT-4o outperformed all other models with an S-CVI/Ave value of 0.943 and S-CVI/UA value of 0.857, both exceeding the acceptable thresholds of 0.90 and 0.80 respectively. This suggests that ChatGPT-4o is a viable model for preparing lecture notes in anatomy. Since it exceeds the established thresholds, this means that ChatGPT-4o is a suitable model to prepare lecture notes in anatomy. ChatGPT-4o-Mini was next with an S-CVI/Ave value of 0.800 with an S-CVI/UA value of 0.714 which achieved a level of endorsement in S-CVI/UA but just missed S-CVI/Ave. Copilot's S-CVI/Ave was 0.486 and S-CVI/UA was 0.286, both well below acceptable thresholds. Gemini showed the least effectiveness with an S-CVI/Ave value of 0.286 and S-CVI/UA of 0.143, falling short of both threshold values. Although ChatGPT-4o-Mini is slightly below the S-CVI/Ave threshold, its performance suggests it may be ready for use in the near future with additional evaluation. However, note that both the Gemini and Copilot models were well-below the acceptable thresholds, indicating substantial work is needed to get both of their capabilities in-line with the type of material expected and suitable for preparing a lecture material in anatomy field.

A recent study reported that ChatGPT-4 provides well-structured and accurate anatomical descriptions, including clinical relevance and structure relationships. Its ability to generate concise chapter summaries and clarify anatomical terminology, even for complex terms, makes it a valuable supplementary tool for students and educators. Additionally, it can create multiple-choice quizzes and matching questions of varying difficulty levels, enhancing its role in educational assessment. Despite these strengths, certain limitations persist. The study also has shown that ChatGPT -4's handling of anatomical variants and their clinical significance is inconsistent, as its responses tend to lack depth unless such variants are systematically classified into types. This suggests that while the AI model performs well with standardized anatomical knowledge, it struggles with more complex and nuanced topics requiring a higher level of interpretative reasoning [12]. Concerns have been raised regarding the accuracy of ChatGPT's anatomical responses. Another study has found that while ChatGPT-4 can provide generally well-structured information, it is prone to factual errors, misinterpretations, and omissions in anatomical details. For example, when responding to fact-based anatomical questions, it has produced incorrect information regarding nerve branches and their functions, highlighting deficiencies in its underlying training data. This raises concerns about the reliability of AI-generated medical content and the necessity of expert validation [15]. Another recent study comparing large language models (LLMs) in medical education has shown that ChatGPT-4 outperforms other AI models, such as ChatGPT-3.5, Copilot, Bard, and Google PaLM, in answering anatomy-related multiple-choice questions and generating clinical scenarios. In an evaluation of chatbot performance on medical multiple-choice questions from a Gross Anatomy course exam database, ChatGPT-4 demonstrated the highest accuracy, answering a significantly more significant proportion of questions correctly than Copilot, ChatGPT-3.5, and other models. The study also assessed the chatbots' ability to generate clinical scenarios and corresponding multiple-choice questions for selected anatomical topics, where ChatGPT-4 again outperformed competing models, followed by Gemini, ChatGPT-3.5, and ChatGPT-3.5-turbo. Despite this relative advantage, the study concluded that LLMs, including ChatGPT-4, have yet to reach a level of maturity where they can fully replace human educators in Gross Anatomy courses, though they can serve as valuable supplementary tools for medical instruction [11]. Furthermore, a comparative study evaluating ChatGPT's performance against undergraduate students in an anatomy course found that ChatGPT outperformed students in a multiple-choice examination. The study, conducted with students from the faculty of health sciences at a university in Turkey, revealed that ChatGPT demonstrated higher accuracy in answering anatomy-related questions than human participants. These results highlight ChatGPT's potential as a practical learning tool for anatomy education, particularly in knowledge-based assessments [16]. A study by Ilgaz and Çelik (2023) investigated the effectiveness of ChatGPT and Google Bard in generating anatomy-related content, finding that while LLMs could generate quizzes and provide general information, their accuracy in article writing remained inconsistent [9]. These findings and the current study suggest that while LLMs offer valuable support in anatomy education, they require further refinement to match the precision of specialized learning tools for anatomy.

To date, no studies have been published assessing the performance of large language models (LLMs) in any formal anatomy licensing and board examinations. However, multiple studies have documented ChatGPT-4's success in various medical examinations that necessitate a substantial understanding of anatomical knowledge. Research by Kung et al. (2023) demonstrated that ChatGPT successfully passed medical licensing exams, indicating its capability to process and convey complex medical knowledge [17]. In a separate study, ChatGPT demonstrated a performance equivalent to that of a third-year medical student sitting for the same exam [18]. Similarly, a recent Korean study reported that ChatGPT passed the Korean General Surgery Board exam [19]. It also has been shown to perform extremely well on the European Exam in Core Cardiology [20]. In another study, it was found that ChatGPT was on par with a first-year plastic surgery resident in terms of the ability to pass the exam [21]. These achievements highlight the potential of

ChatGPT to access and convey the essential knowledge needed to excel in such difficult and niche examinations. These research studies, along with the current study, demonstrate that ChatGPT offers medical educators accurate and up-to-date information while also serving as a practical and valuable learning tool for students. For students who need to soak up huge amounts of information a short time, LLMs may offer instant access to key medical content, research, and clinical guidelines, simplifying the learning process.

ChatGPT-4o may serve as a useful tool in medical education, particularly in areas where access to new medical textbooks or academic information is limited due to financial or infrastructural constraints. A study by Tung and Dong found that Malaysian medical students are not only aware of AI but also show a strong interest in learning more about its applications [22]. Similarly, Buabbas et al. reported a positive attitude among students toward AI in medical education, with many expressing the belief that it can effectively support both teaching and learning processes [23].

The findings of this study highlight ChatGPT-4o's acceptable performance in delivering accurate, relevant, clear, and comprehensive information, firmly establishing its potential as a valuable tool in medical education in anatomy field. However, similar results have not been consistently observed with other models, underscoring the variability in their capabilities. While ChatGPT-4o holds significant promise for transforming anatomy education, concerns about its reliability and potential effects on students' critical thinking abilities remain. To address these challenges, it is essential to develop robust systems for verifying the accuracy of its outputs and to integrate ChatGPT-4o as a complementary tool that enhances, rather than replaces, traditional teaching methods.

Limitations

This study has several limitations. Firstly, the findings are confined to the timeframe during which the study was conducted, as the LLMs generating the content are continually updated and improved. Secondly, while the CVI analysis adhered to the internationally accepted number of experts, it may still be considered relatively limited. Future studies involving the evaluation of LLM-generated content by a broader and more diverse panel of experts are likely to provide more comprehensive and enlightening results.

CONCLUSION

The findings of this study indicate that ChatGPT-4o demonstrated the highest performance in terms of thematic coverage and content validity index, meeting acceptable scientific thresholds for validity and accuracy. However, other models, such as ChatGPT-4o-mini, Gemini, and Copilot, exhibited significantly lower performance metrics. These findings are based on quantitative analysis using the Content Validity Index (CVI), which is a widely accepted statistical method for evaluating content reliability, as well as thematic analysis, which—although qualitative—includes quantifiable aspects such as frequency counts and inter-rater agreement. Based on these findings, we recommend that anatomy educators and medical students consider using this model as a complementary content-generation tool for anatomy education. However, other models such as ChatGPT-4o-mini, Gemini, and Copilot require significant improvement before they can be reliably used to produce anatomy lecture notes.

Conflict of Interest: None declared.

Funding: The authors received no financial support for the research.

Informed Consent: Since this study utilized publicly available data/secondary data/literature review, informed consent was not applicable.

Authors' Contribution: Conception: FO, BK, FTK; Design: FO Supervision: FO; Materials: FO, BK, FTK; Data Collection and/or Processing: FO, BK, FTK; Analysis and/or Interpretation: FO; Literature review: FO; Writing: FO; Critical Review: FO, BK. All authors approved the final version of the manuscript.

Ethical Approval: Since this study utilized publicly available data/secondary data/literature review, ethical approval was not applicable.

REFERENCES

[1] Ghassemi M, Birhane A, Bilal M, Kankaria S, Malone C, Mollick E, Tustumi F (2023) ChatGPT one year on: who is using it, how and why?. Nature. 624(7990):39-41. https://doi.org/10.1038/d41586-023-03798-6

[2] Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations,

prospects, and ethical considerations. Front Artif Intell. 6:1169595. https://doi.org/10.3389/frai.2023.1169595

[3] Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, Rios CJG, Godoy CAR, Quezada KLTP, Gutierrez-Arratia JS, Flores-Cohalia JA (2023) Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 20:30. https://doi.org/10.3352/jeehp.2023.20.30

[4] Ornstein J (1955) Mechanical Translation: New Challenge to Communication. Science. 122(3173):745-748. https://doi.org/10.1126/science.122.3173.745

[5] Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, Cuocolo R, Cannella R, Koçak B (2024) Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 30(2):80-90. https://doi.org/10.4274/dir.2023.232417

[6] Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen D (2024) Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard. JMIR Med Educ. 10:e51523. https://doi.org/10.2196/51523

[7] Sallam M (2023) ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 11(6):887. https://doi.org/10.3390/healthcare11060887

[8] Lee H (2024) The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 17(5):926-931. https://doi.org/10.1002/ase.2270

[9] Ilgaz HB, Çelik Z (2023) The Significance of Artificial Intelligence Platforms in Anatomy Education: An Experience With ChatGPT and Google Bard. Cureus. 15(9):e45301. https://doi.org/10.7759/cureus.45301

[10] Arun G, Perumal V, Urias FPJB, Ler YE, Tan BWT, Vallabhajosyula R, Tan E, Ng O, Ng KB, Mogali SR (2024) ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study. Anat Sci Educ. 17(7):1396–1405. https://doi.org/10.1002/ase.2502

[11] Mavrych V, Ganguly P, Bolgova O (2025) Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. Clin Anat. 38(2):200–210. https://doi.org/10.1002/ca.24244

[12] Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, Piagkou M (2023) The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 45(10):1321–1329. https://doi.org/10.1007/s00276-023-03229-1

[13] Moore KL, Dalley AF (2018) Clinically oriented anatomy, 8th edn. Wolters kluwer, India.

[14] Shi J, Mo X, Sun Z (2012) [Content validity index in scale development]. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 37(2):152-155. https://doi.org/10.3969/j.issn.1672-7347.2012.02.007

[15] Mogali SR (2024) Initial impressions of ChatGPT for anatomy education. Anat Sci Educ. 17(2):444–447. https://doi.org/10.1002/ase.2261

[16] Talan T, Kalınkara Y (2023) The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course. UYBISBBD. 7(1):33-40. https://doi.org/10.33461/uybisbbd.1244777

[17] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198

[18] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D (2023) 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. 9:e45312. https://doi.org/10.2196/45312

[19] Oh N, Choi GS, Lee WY (2023) ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res. 104(5):269-273. https://doi.org/10.4174/astr.2023.104.5.269

[20] Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S (2023) ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?. Eur Heart J Digit Health. 4(3):279-281.

https://doi.org/10.1093/ehjdh/ztad029

[21] Humar P, Asaad M, Bengur FB, Nguyen V (2023) ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination. Aesthet Surg J. 43(12):Np1085-np1089. https://doi.org/10.1093/asj/sjad130

[22] Tung AYZ, Dong LW (2023) Malaysian Medical Students' Attitudes and Readiness Toward AI (Artificial Intelligence): A Cross-Sectional Study. J Med Educ Curric Dev. 10:23821205231201164. https://doi.org/10.1177/23821205231201164

[23] Buabbas AJ, Miskin B, Alnaqi AA, Ayed AK, Shehab AA, Syed-Abdul S, Uddin M (2023) Investigating Students' Perceptions towards Artificial Intelligence in Medical Education. Healthcare (Basel). 11(9):1298. https://doi.org/10.3390/healthcare11091298

How to Cite;

Ok F, Karip B, Temizsoy Korkmaz F (2025) PEvaluating the Performance of Large Language Models in Anatomy Education. Eur J Ther. 31(1):35-43. https://doi.org/10.58600/eurjther2611