How Well Did ChatGPT Perform in Answering Questions on Different Topics in Gross Anatomy?

Olena Bolgova, Inna Shypilova, Larysa Sankova, and Volodymyr Mavrych

ABSTRACT

The burgeoning interest in leveraging ChatGPT within the medical field underscores the necessity for a comprehensive understanding of its capabilities and limitations, particularly in the context of medical assessments and examinations. The model possesses a unique aptitude for addressing queries related to medical student exams, thereby serving as an invaluable resource for academic support. Its advanced natural language processing capabilities empower it to comprehend the intricacies of medical terminology, enabling it to provide nuanced and contextually relevant responses. This study aimed to quantitatively evaluate ChatGPT performance in answering Multiple Choice Questions (MCQs) related to the different topics in Gross Anatomy course for medical students.

The research conducted for this study was focused on a comprehensive examination of ChatGPT (GPT-3.5) capabilities in answering 325 MCQs designed in USMLE style, arranged in 7 different sets related to specific topics. These questions were selected from Gross Anatomy course exam database for medical students and reviewed by three independent experts. The results of 5 successive attempts to answer each set of questions by ChatGPT were evaluated based on accuracy, relevance, and comprehensiveness.

The ChatGPT provided accurate answers to 44.1% ± 8.2% of questions. Accordingly, to our data, ChatGPT is answering much better on MCQs from Back material (58.4%), following Head and Neck (48.8%) and Pelvis (45.6%), and performing not so well in questions of Thorax (37.6%) and Upper limb (36.4%). ChatGPT is struggling in answering questions about blood supply and innervation of the specific organs.

ChatGPT stands out as a promising and interactive educational tool, particularly for students engaged in the study of anatomy. Its distinctive ability to not only provide informative responses but also engage students in a conversational manner is highly commendable. This quality has the potential to enhance student engagement and foster curiosity, creating a dynamic learning experience. However, it’s crucial to acknowledge that ChatGPT’s current level of comprehension and interpretative abilities may not meet the demanding standards required for practical applications in the medical education domain. Its performance in challenging examinations like medical college exams and health licensing exams might need to catch up to expectations.

Keywords: Artificial intelligence, ChatGPT, Gross Anatomy, Medical education.

1. INTRODUCTION

Chat Generative Pretrained Transformer (ChatGPT) represents an advanced natural language processing model encompassing a 175-billion-parameter architecture and leveraging deep learning algorithms and training methodologies on extensive datasets. Developed by OpenAI, ChatGPT belongs to the family of generative pre-training transformer (GPT) models, standing out as one of the most expansive publicly available language models [1], [2]. Its main job is to produce responses to various natural language inputs similar to those of a human, a capability ascribed to its deep learning base [3].

Despite the extensive use of Artificial Intelligence (AI) in domains such as customer support and data management,
its integration into healthcare and medical research sectors has encountered notable limitations [4]. The potential applications of ChatGPT within the medical domain are multifaceted, ranging from identifying research topics to aiding professionals in clinical and laboratory diagnosis [5]. However, implementing ChatGPT in medical contexts is challenging, encompassing concerns about credibility, plagiarism, and ethical considerations [6]–[8].

One significant application of ChatGPT in the medical field lies in developing virtual assistants designed to assist patients in managing their health. These virtual assistants can automate the summarization of patient interactions and medical histories, streamlining the medical recordkeeping process for healthcare practitioners [4]. Additionally, ChatGPT exhibits the potential to aid patients in managing their medications by providing reminders, dosage instructions, and information about potential side effects, drug interactions, and other relevant considerations [9].

ChatGPT’s role in the medical domain extends beyond mere conversational interactions, as it has been specifically trained on conversational prompts to foster dialogic output [10], [11]. Despite demonstrating potential, language models (LLMs) such as ChatGPT have faced challenges in testing clinical knowledge through generative question-answering tasks [12], [13].

The impact of ChatGPT as a Natural Language Processing (NLP) is profound, with its continuous evolution expected to significantly influence the landscape of conversational AI across various sectors [14]. In the medical field, workers are increasingly incorporating AI, including ChatGPT, to enhance efficiency in areas such as diagnosis, medical imaging analysis, predictive models, and personalized medicine [15]. ChatGPT holds promise for improving these applications by amalgamating medical knowledge with conversational capabilities.

ChatGPT’s utilization extends to medical licensing exams in the United States, demonstrating notable advancements in natural language processing, particularly in answering medical questions [10]. Its ability to convey logical and contextual information underscores its potential as a valuable medical education and learning support tool.

As ChatGPT’s application in medical examinations gains attention, there is a need to assess its accuracy in solving medical questions, given its recent development as an AI chatbot (OpenAI). ChatGPT offers personalized learning materials in the education sector and the capacity to address queries related to medical student exams [16]. Its advanced natural language processing capabilities enhance the learning experience, making it more efficient and engaging for students.

ChatGPT’s contributions to medical education encompass diverse functionalities, including evaluating essays and papers and analyzing sentence structure, vocabulary, grammar, and clarity [13]. Furthermore, it can generate exercises, quizzes, and scenarios for classroom use, aiding in practice and assessment. The model’s ability to write basic medical reports assists students in identifying areas for improvement and deepening their understanding of complex medical concepts. Additionally, ChatGPT’s translation, explanation, and summarization capabilities facilitate comprehension of intricate learning materials [17].

For ChatGPT to be effective in these educational applications, it must demonstrate performance comparable to human experts in medical knowledge and reasoning tests, instilling confidence in users regarding the reliability of its responses [1]. ChatGPT is a versatile tool that can revolutionize various facets of society, from business development to medicine, education, research, coding, entertainment, and art [14], [18].

Compared to questions crafted by university professors, a compelling investigation was conducted to assess the efficacy of multiple-choice questions (MCQs) generated by ChatGPT for medical graduate examinations. ChatGPT demonstrated remarkable efficiency by producing 50 MCQs within 20 minutes, starkly contrasting to human examiners who expended over 211 minutes for an equivalent set of questions [19]. This discrepancy in time investment signifies the potential expeditiousness offered by ChatGPT in item development for medical assessments.

An evaluation conducted by impartial experts revealed that, notwithstanding a slight dip in the domain of relevance, ChatGPT’s generated questions exhibited a level of quality comparable to those crafted by human counterparts. ChatGPT’s questions showcased a broader spectrum of scores, contrasting with the more consistent scoring observed in questions devised by humans. In summation, while ChatGPT holds promise in the generation of high-quality MCQs for medical graduate examinations, its performance, particularly in terms of relevance, invites further scrutiny. Nevertheless, it positions itself as an efficacious next-generation technique with the potential to address prevailing challenges in item development, presenting an economically viable solution for the future of medical assessment [20].

Explorations into ChatGPT’s aptitude in guiding medical students within the domain of anatomy education and research have yielded insightful findings. Interactions involving queries directed at ChatGPT assessed its accuracy, relevance, and comprehensiveness in delivering anatomical information. ChatGPT demonstrated proficiency in providing accurate anatomical descriptions imbued with clinical significance and structural relationships. Furthermore, it exhibited adeptness in furnishing summaries and offering terminology assistance. Nonetheless, improvements are deemed necessary, particularly in systematically classifying responses to anatomical variants [21].

Recent publications have underscored the positive outcomes arising from the utilization of ChatGPT in answering multiple-choice questions, signaling potential transformative impacts on the educational system. In examining the accuracy and consistency of responses, ChatGPT-3.5 outperformed Google Bard when responding to lung cancer prevention, screening, and radiology terminology queries. ChatGPT-3.5 achieved a correctness rate of 70.8%, while Google Bard attained a rate of 51.7% [22]. These findings emphasize the proficiency of ChatGPT-3.5 in delivering accurate and reliable responses in the medical knowledge domain.
ChatGPT’s competence in addressing typical patient queries about optic disc drusen and total hip arthroplasty has been demonstrated, providing generally relevant information. However, a caveat arises as specific responses, particularly those about treatment and prognosis, were found to need to be more accurate. Such inaccuracies pose potential harm and underscore the importance of exercising caution when relying exclusively on ChatGPT for patient information and underscores the imperative of integrating human oversight in the utilization of ChatGPT in medical contexts to ensure the provision of accurate and safe information to patients [23].

Even though ChatGPT was only recently introduced, there are many publications on this topic, but only some have statistical analysis included. The main objectives of our research were to develop an algorithm for the quantitative analysis of the Chatbot’s ability to answer MCQ tests and check its effectiveness in medical education, specifically in the different topics of the Gross Anatomy course for medical students.

2. MATERIALS AND METHODS

The research conducted for this study was focused on a comprehensive examination of ChatGPT capabilities in answering 325 USMLE-style multiple-choice questions (MCQs). The questions were randomly selected from the Gross Anatomy course exam database for medical students and reviewed by three independent experts. No questions with images were included in this study. The selected questions had different levels of difficulty. They have been subdivided into the following seven sets: Abdomen (AB), Back (BK), Head and Neck (HN), Lower Limb (LL), Pelvis (PL), Thorax (TH), and Upper limb (UL). Each set included 50 MCQs, except for the back material, which had only 25 questions. Since all questions were created in 2020, we avoided the lack of real-time information limitation for ChatGPT. GPT-3.5 knowledge is based on text data up to September 2021 since it cannot access real-time information on events that occurred after that date.

The results of 5 successive attempts by ChatGPT to answer these questionary sets were evaluated based on accuracy, relevance, and comprehensiveness. Each ChatGPT attempt’s data was recorded and compared with all previous attempts, finding the percentage of repeated and correct answers among them.

Seven sets of random answers were generated and analyzed for the same MCQ sets utilizing the RAND () function in Microsoft Excel (Microsoft®365) to compare the results of ChatGPT performance with random guessing. Statistica 13.5.0.17 (TIBCO® Statistica™) was used to analyze the data’s basic statistics and compare results among the different MCQ sets.

3. RESULTS

According to our data, ChatGPT accurately answered 44.1% ± 8.2% of 325 multiple-choice questions (MCQs) from various topics from Gross Anatomy course, which is much better than random responses (19.0% ± 6.4%) for the same set of the questions. However, there is a significant variation in ChatGPT answers among the different topics: the best result was recorded for the Back questions, then for Head and Neck, followed by Pelvis, Abdomen, Lower limb, and Thorax, and the lowest were results for the Upper limb material (Fig. 1). After that the detailed evaluation of results received for each topic (set of the questions) was done.

3.1. Back MCQs

The results of five attempts of ChatGPT to answer the set of 25 BK MCQs had shown 58.4% ± 4.6% correct answers. The random generation of answers for the same set of MCQs gives 18.4% ± 9.2% correct answers only. The first attempt of ChatGPT was the most successful, with 64% correct answers. The results of the next four successive attempts fluctuated in the range of 52%–60%. The coincidence of answers with the previous generations was in the interval of 64%–72%, and among them, the coincidence of correct answers was 48%–52% (Table I).

12 questions (48%) were answered correctly across all five attempts and considered a solid knowledge area for ChatGPT. The item analysis indicated that these MCQs were about muscles of the back, spine, embryology, and CSF. They were recall questions. ChatGPT did not show good results in answering more comprehensive questions.
TABLE II: ChatGPT % of Correct Answers, Coincidence with a Previous Attempt, and Coincidence of Correct Answers with a Previous Attempt for 5 Attempts from HN MCQs

<table>
<thead>
<tr>
<th>Attempt number:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct answers</td>
<td>50</td>
<td>52</td>
<td>48</td>
<td>46</td>
<td>48</td>
</tr>
<tr>
<td>Coincidence with 1</td>
<td>66</td>
<td>76</td>
<td>58</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Coincidence corrects with 1</td>
<td>42</td>
<td>42</td>
<td>36</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Coincidence with 2</td>
<td>60</td>
<td>64</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coincidence corrects with 2</td>
<td>36</td>
<td>36</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coincidence with 3</td>
<td>64</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coincidence corrects with 3</td>
<td>36</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coincidence with 4</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coincidence corrects with 4</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2. Head and Neck MCQs

After five attempts, ChatGPT generated 48.8% ± 2.3% accurate answers for the set of 50 HN MCQs. The random generator yielded only 21.6% ± 8.0% accurate answers for the identical multiple-choice questions. The initial attempt at ChatGPT was not very successful; just 50% of the responses were right. The following four tries yielded outcomes ranging from 46% to 52%. The answers coincided with those of the preceding attempts in a range of 58% to 76%; among them, the coincidence of correct answers was 60% to 76% (Table II).

We considered the 13 questions (26%) a strong knowledge area for ChatGPT because they were answered correctly on each of the five tries. The item analysis indicated that these MCQs were about bones of the skull, nose, tonsils, larynx, pharynx, embryology, and pituitary gland. They were recall questions. ChatGPT did not do well in answering questions about blood vessels, branches of the spinal nerves, and topographic regions of the back.

3.3. Pelvis MCQs

The results of five attempts of ChatGPT to answer the set of 50 PL MCQs had shown 45.6% ± 5.0% correct answers. The random generation of answers for the same set of MCQs gives 19.2% ± 8.8% correct answers only. The first ChatGPT attempt was not the most successful, with only 42% of the answers being correct. The results of the next four successive attempts fluctuated in the range of 42%–54%. The coincidence of answers with the previous generations was in the interval of 56%–68%, and among them, the coincidence of correct answers was 34%–40% (Table III).

14 questions (28%) were answered correctly across all five attempts and considered a solid knowledge area for ChatGPT. The item analysis indicated that these MCQs were about the bony pelvis, uterus, prostate gland, urethral sphincters, and pelvic lymphatic nodes. ChatGPT did not do well in answering comprehensive questions about external genital organs, blood vessels, innervation of the pelvic viscera, muscles, fascia, and spaces of the perineum.

3.4. Abdomen MCQs

After five tries, ChatGPT’s scores for the 50 AB MCQs revealed 41.6% ± 4.6% accurate answers. The random generation of answers for the same set of MCQs gives 16.4% ± 3.3% correct answers only. With 46% of the answers correctly, the first ChatGPT attempt was one the most successful. The next four tries were 38% to 46% of the outcomes. 56%–68% of the answers were similar to those from earlier generations; among them, between 30% and 34% were right (Table IV).

12 questions (24%) were answered correctly across all five attempts and considered a solid knowledge area for ChatGPT. The item analysis indicated that these MCQs were about the inguinal region, diaphragm, pancreas, liver, and small intestine. ChatGPT did not answer more comprehensive questions about abdominal blood vessels, innervation of the abdominal viscera, embryology, large intestine, and peritoneum.

3.5. Lower Limb MCQs

The results of five attempts of ChatGPT to answer the set of 50 LL MCQs had shown 40.0% ± 5.1% correct answers. The random generation of answers for the same set of MCQs gives 18.0% ± 5.1% correct answers only. The first ChatGPT attempt was 44% of the correct answers. The results of the next four successive attempts were in the 32%–44% range. The coincidence of answers with the previous generations was in the interval of 42%–58%, and among them, the coincidence of correct answers was 26%–30% (Table V).

9 questions (18%) were answered correctly across all five attempts and considered a solid knowledge area for
3.6. Thorax MCQs

The results of five attempts of ChatGPT to answer the set of 50 TH MCQs had shown 37.6% ± 3.3% correct answers. The random generation of answers for the same set of MCQs gives 18.8% ± 4.8% correct answers. The first ChatGPT attempt gave 34% of the correct answers. The results of the next four successive attempts were more successful in the 36%–42% range. The coincidence of answers with the previous generations was in the interval of 44%–62%; among them, the coincidence of correct answers was 16%–26% (Table VI).

Only 6 questions (12%) were answered correctly across all five attempts and considered a solid knowledge area for ChatGPT. The item analysis indicated these MCQs were about the upper limb nerve and muscle injuries. ChatGPT did not answer questions about the major branches of the brachial plexus, blood vessels, muscles, facia, joints of the upper limb, and lymphatic nodes.

3.7. Upper Limb MCQs

The results of five attempts of ChatGPT to answer the set of 50 UL MCQs had shown 36.4% ± 3.8% correct answers. The random generation of answers for the same set of MCQs gives 20.8% ± 5.8% correct answers. The first ChatGPT attempt gave only 30% of the correct answers. The results of the next four successive attempts were more successful in the 36%–40% range. The coincidence of answers with the previous generations was in the interval of 44%–62%; among them, the coincidence of correct answers was 16%–26% (Table VII).

Only 6 questions (12%) were answered correctly across all five attempts and considered a solid knowledge area for ChatGPT. The item analysis indicated these MCQs were about the upper limb nerve and muscle injuries. ChatGPT did not answer questions about the major branches of the brachial plexus, blood vessels, muscles, facia, joints of the upper limb, and lymphatic nodes.

4. Discussion

In our investigation, it was observed that ChatGPT exhibits a commendable capability to respond accurately to Multiple-Choice Questions within the context of the Gross Anatomy course, achieving an average accuracy rate of 44.1%, which is significantly above the random responses to the same questions (19%). Our results align notably with a recent study evaluating ChatGPT’s performance in answering questions from the US Medical Licensing Exams, where accuracy ranged from 42% to 64.4%, surpassing other models like Instruct GPT [1]. When evaluated against the Chinese National Medical Licensing Examination in both 2021 and 2022, ChatGPT’s scores fell short of meeting the passing requirements. In comparison to the national pass rates of 50% in 2021 and 55% in 2022, ChatGPT’s performance currently falls below the standard achieved by medical students who have undergone traditional 5-year medical education in a medical school [15].

Previous research within the realm of medical question answering has often been tailored to specific tasks, aiming to enhance model performance at the expense of generalizability. For instance, [24] achieved a 68.1% accuracy in their model, which specialized in answering yes-or-no questions based on information found in PubMed-available abstracts.

Controversy, a study evaluating ChatGPT’s performance in answering questions from the 2022 Brazilian National Examination for Medical Degree Revalidation (Revalida) revealed an impressive accuracy of 87.7%, with no significant differences observed across medical topics [25]. Answering questions from the European Board of Ophthalmology, ChatGPT demonstrated remarkable proficiency, achieving a stellar 91% success rate and
showing an in-depth understanding and application of ophthalmology knowledge [26]. ChatGPT is much more efficient in providing knowledge, management, and emotional support for patients with cirrhosis and hepatocellular carcinoma. While it demonstrated commendable knowledge retention for cirrhosis (79.1% correct) and hepatocellular carcinoma (74.0% correct), limitations were identified in responses lacking comprehensiveness, particularly in diagnosis and prevention aspects [27].

However, efforts towards more generalizable models encountered more significant challenges, as evidenced by different studies achieving lower accuracies, such as 36.7% and 29% on datasets derived from Chinese medical licensing exams and USMLE Step 1 and Step 2 questions, respectively [1], [15]. This observation is consistent with our data, wherein only 9%-48% of the questions were correctly answered across all five attempts, particularly highlighting challenges in responding to more comprehensive queries beyond simple recall questions.

Speaking about different medical subjects, ChatGPT demonstrates enhanced performance in physiology, head and neck surgery, and biochemistry, particularly when confronted with diverse question types beyond MCQs. A study underscores ChatGPT’s effectiveness in addressing reasoning questions related to core physiology concepts, achieving an approximately 74% correct response rate [28]. In head and neck surgery, ChatGPT exhibited notable success with correct responses in closed-ended questions (84.7%) and accurate diagnoses in clinical scenarios (81.7%). However, some aspects, such as the completeness of proposed procedures and the quality of bibliographic references, were noted as areas for improvement [29].

Another study focusing on medical biochemistry revealed that ChatGPT’s median score of 80% indicates the need for ongoing training and development for higher-order questions [30].

Our investigation has yielded intriguing insights into ChatGPT’s performance across distinct anatomical topics, particularly in its responsiveness to MCQs. Notably, the model demonstrated a high level of proficiency in addressing questions related to the Back material, achieving an impressive accuracy rate of 58.4%. This robust performance underscores ChatGPT’s commendable grasp of anatomical concepts associated with the vertebral column and posterior structures. Moving on to the Head and Neck, ChatGPT maintained a solid performance, securing an accuracy rate of 48.8%. This indicates the model’s proficiency in tackling queries related to the intricate anatomy of the head and neck region. Similarly, in addressing questions concerning the Pelvis, ChatGPT exhibited competence, with an accuracy rate of 45.6%. This suggests a notable understanding of pelvic anatomy and associated structures.

However, our findings reveal a comparative decline in ChatGPT’s performance when confronted with questions related to the Thorax, where it registered an accuracy rate of 37.6%. This lower accuracy suggests that the model encounters challenges in accurately responding to queries about thoracic anatomy, potentially reflecting the complexity of this anatomical region. Furthermore, ChatGPT demonstrated a relatively lower accuracy rate of 36.4% in Upper Limb anatomy. This indicates a notable area for improvement, signifying that the model faces challenges when addressing questions related to the anatomy of the upper extremities.

Our observation is of particular interest that ChatGPT encounters difficulties in responding to questions involving the blood supply and innervation of specific organs. This suggests a nuanced challenge within the model’s comprehension and retrieval of information related to vascular and neural structures within the human body.

These nuanced variations in ChatGPT’s performance across different anatomical topics underscore the intricacies of anatomical knowledge representation within the model. While excelling in certain areas, the model faces challenges in others, revealing potential avenues for refinement and improvement. As natural language processing and medical AI progress, addressing these specific challenges could further enhance ChatGPT’s utility and reliability in medical education and anatomical understanding.

5. Conclusion
ChatGPT stands out as a promising and interactive educational tool, particularly for students engaged in the study of anatomy. Its distinctive ability to provide informative responses and conversationally engage students is highly commendable. This quality can enhance student engagement and foster curiosity, creating a dynamic learning experience.

However, it is crucial to acknowledge that ChatGPT’s current comprehension and interpretative capabilities may not meet the demanding standards required for practical applications in the medical education domain. In particular, its performance in challenging examinations like medical school and health licensing exams might need to meet expectations.

Nevertheless, there is optimism about the future development of ChatGPT. ChatGPT’s knowledge base and interpretative capabilities are expected to evolve rapidly due to recent advancements in deep learning. This evolution could pave the way for the integration of ChatGPT into various aspects of medical and health education. Both educators and students are encouraged to stay abreast of these developments, recognizing the potential benefits and opportunities that may arise from incorporating this AI platform into the educational landscape.

As AI advances, the ongoing refinement of ChatGPT’s capabilities holds promise for its broader utility in medical education. The prospect of leveraging ChatGPT as a supplementary resource, aiding students in understanding complex anatomical concepts and beyond, is an exciting avenue to explore. This could potentially revolutionize the educational landscape, offering a novel and interactive approach to learning that complements traditional teaching methods.

In conclusion, while recognizing the current limitations, it is crucial to view ChatGPT as an evolving educational resource with the potential to play a significant role in the future of medical and health education. As the technology evolves...
The authors received no funding for this study.

Conflict of Interest
Authors declare that they do not have any conflict of interest.

References