information

MDPI

Article

A Multi-Chatbot Analysis: Strengths and Weaknesses in Neuroanatomy Learning

Alessandro Naim [1,2,†], Sara Naim [1,2,†] and Daniele Saverino [3,4,*]

[1] IRCCS 'G. Gaslini' Institute, 16147 Genova, Italy; alessandronaim@gaslini.org (A.N.); 4201144@studenti.unige.it (S.N.)
[2] Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, University of Genoa, 16132 Genova, Italy
[3] Department of Experimental Medicine, Section of Human Anatomy, University of Genoa, 16132 Genova, Italy
[4] IRCCS Ospedale Policlinico San Martino, 16132 Genova, Italy
* Correspondence: daniele.saverino@unige.it
† These authors contributed equally to this work.

Abstract

Background: The expanding interest in chatbots within the medical domain underscores the imperative for a comprehensive understanding of their capabilities and limitations, particularly in the context of anatomical education. Chatbots possess the potential to comprehend intricate anatomical concepts, deliver both advanced and contextually relevant information, and serve as a valuable resource for medical students and educators. This study aimed to evaluate the proficiency and constraints of chatbots in the domain of neuroanatomy. **Methods:** We developed 30 questions and administered them to ChatGPT-4, Google Gemini, Microsoft Copilot, and Perplexity.ai in their open versions. Questions were collaboratively constructed by the research team, selected through a semi-randomized process within the domain of neuroanatomy. Chatbots' responses were evaluated in a blinded manner for validity and appropriateness, utilizing a 5-point Likert scale. **Results:** The highest observed performance among the evaluated chatbots was exhibited by ChatGPT-4 and Perplexity.ai, which achieved scores of 4.6 ± 0.5 and 4.5 ± 0.5, respectively. Microsoft Copilot (4.4 ± 0.5) and Google Gemini (4.1 ± 1.0) followed. The least successful performance was observed in the task of generating a neuroanatomical structure: only Microsoft Copilot attempted to fulfil the request, albeit with a dramatically flawed outcome. Conversely, Google Gemini and Perplexity.ai provided web links to anatomical illustrations. **Conclusions:** Despite technological advancements, AI models have not yet reached a level of sophistication sufficient to entirely supplant the role of educators or facilitators in a neuroanatomy course; however, they can serve as valuable adjunct tools for medical educators and students when utilized with careful consideration.

Keywords: neuroanatomy; artificial intelligence; chatbots; medical education

Academic Editors: Joao Carlos Lopes Batista, Dimitris Apostolou and Anabela Mesquita

Received: 9 April 2026
Revised: 29 April 2026
Accepted: 7 May 2026
Published: 13 May 2026

1. Introduction

Neuroanatomy has historically been regarded as a keystone of medical education, providing the essential foundation for understanding the structure and function of the nervous system [1,2]. Mastery of neuroanatomical knowledge is critical for accurate diagnosis, interpretation of imaging studies, and planning surgical interventions, particularly in neurology, neurosurgery, and radiology [3]. Despite its importance, neuroanatomy is

often perceived as a challenging subject due to its complexity and the need for spatial visualization of intricate structures [4].

In recent years, artificial intelligence (AI) has emerged as a transformative force in medical education, driven by the development of large language models (LLMs) capable of generating human-like responses to complex queries [5,6]. AI-based chatbots such as ChatGPT-4, Google Gemini, Microsoft Copilot, and Perplexity.ai leverage vast datasets to provide rapid, contextually relevant information, offering opportunities for personalized learning, interactive question-answering, and real-time feedback [7,8]. These tools have been explored in various educational contexts, including anatomy, where they promise to complement traditional teaching methods and enhance student engagement [9,10]. The integration of AI technologies has the potential to significantly alleviate the substantial workload associated with anatomy departments worldwide, where cadaveric dissection would be a cornerstone of teaching and learning. However, it is crucial to exercise caution and rigorously evaluate the credibility of information provided by AI-powered tools [3,6,7,9–17].

Recent comparative studies have highlighted the strong performance of AI-based large language models (LLMs) in anatomy-related multiple-choice questions, reporting accuracy rates exceeding 95% across platforms such as Google Gemini, ChatGPT-4, and DeepSeek, while also emphasizing the need for expert oversight to mitigate occasional errors [17].

However, significant limitations persist. Current AI models struggle to generate accurate visual content, a critical component of anatomy education, and may produce incomplete or inaccurate textual responses [3,6,7,9–17]. These shortcomings raise concerns about their reliability as standalone educational tools. Ethical considerations, including academic integrity and the risk of over-reliance on AI, further underscore the need for cautious implementation [18,19]. Previous studies have reported variable performance of AI platforms in anatomy-related tasks, with accuracy rates ranging from moderate to high depending on question type and complexity [3,6,7,9–17]. While some investigations demonstrated ChatGPT-4's ability to outperform students in multiple-choice assessments [20], others highlighted its inconsistency and inability to provide detailed explanations or accurate diagrams [5,10,21].

While previous studies have examined chatbot performance in anatomy, the present work specifically extends this literature by focusing on neuroanatomy and by systematically comparing multiple widely used chatbots under identical, blinded evaluation conditions. Against this backdrop, the present study aims to evaluate the strengths and limitations of four widely used AI-based chatbots in answering neuroanatomy questions of varying formats and complexity.

2. Materials and Methods

2.1. Study Design and Chatbot Selection

This study employed a comparative evaluation of four freely accessible AI chatbots, ChatGPT-4, Google Gemini, Microsoft Copilot, and Perplexity.ai, to assess their performance in answering neuroanatomy-related questions. Each chatbot was tested using the same set of prompts to ensure consistency. The questions were designed to reflect core neuroanatomy topics typically covered in undergraduate medical curricula, including cranial nerves, sensory and motor pathways, cortical organization, and cerebellar anatomy. The primary aim was comparative assessment across chatbots, rather than formal curriculum validation. Question difficulty was intentionally heterogeneous, spanning recall-based, integrative, and spatially demanding tasks.

2.2. Question Development

A total of 30 questions were designed by a team of neuroanatomy experts. These questions were semi-randomly selected to cover diverse aspects of neuroanatomy and were categorized into four formats: multiple-choice questions (MCQs), questions Q1–Q10; fill-in-the-blank questions, questions Q11–Q20; descriptive questions, questions Q21–Q30, requiring concise and accurate explanations; illustration-based prompts–Embedded within descriptive questions (e.g., Q21, Q22, Q23, Q25, Q28), requesting diagrams or visual aids (Table 1). The format of questions (multiple-choice, fill-in-the-blank, descriptive, illustration-based) was fixed by design, while specific anatomical topics within each format were selected semi-randomly to maximize content diversity and avoid redundancy across the neuroanatomy domain. Finally, illustration-based requests were embedded within descriptive questions. The findings related to image generation are presented as qualitative observations rather than isolated quantitative outcomes. Specifically, the sampling frame consisted of core neuroanatomy topics commonly covered in undergraduate medical curricula (e.g., cranial nerves, sensory and motor pathways, cortical organization, cerebellar anatomy). Within each predefined question format, specific anatomical topics were selected in a semi-random manner from this pool to ensure broad coverage while avoiding topic duplication. The question formats and total number of items were fixed a priori and were not subject to randomization.

Table 1. Questions asked to Chatbots.

Question Number	Questions
1	Most neurons in the central nervous system belong to the category of: a. pseudounipolar; b. bipolar; c. multipolar; d. unipolar
2	Which of the following is a mechanoreceptor for touch? a. Meissner's corpuscles; b. Golgi tendon organs; c. muscle spindles; d. free nerve endings
3	Where are muscle spindles located? a. in smooth muscles; b. in striated muscles; c. in the heart; d. a+b
4	Match the following nerves to their correct spinal nerve root origin (L1, L2, L3, or L4): a. Ilioinguinal nerve; b. Genitofemoral nerve; c. Lateral femoral cutaneous nerve; d. Iliohypogastric nerve
5	The trapezoid body is formed by fibers coming from: a. the red nucleus; b. The trapezoid body is formed by fibers coming from the facial nerve nucleus; c. The trapezoid body is formed by fibers coming from the ventral cochlear nucleus; d. The trapezoid body is formed by fibers coming from the vestibular nucleus; e. The trapezoid body is formed by fibers coming from the inferior olivary nucleus.
6	The inferior horn of the lateral ventricle is located in the lobe: a: frontal; b: parietal; c: temporal, d: occipital; e: all wrong
7	Superiorly, the superior colliculi are in relation with: a. the pituitary gland; b. the cerebellar tonsil; c. the mammillary bodies; d. the tuber cinereum; e. the pineal gland
8	Which of the following statements about the trigeminal nerve (V pair) is correct? a. Sensory afferents derive from the I and II branches (ophthalmic and maxillary) and are directed to the principal sensory nucleus (mesencephalic); b. Sensory afferents derive from the 3 branches (ophthalmic, maxillary, mandibular) and are directed only to the principal sensory nucleus (bulbar); c. Sensory afferents derive from the 3 branches and are directed to the principal sensory nucleus (pontine), the mesencephalic nucleus and the nucleus of the descending root; d. The V pair of cranial nerves has no sensory afferents; e. None of the answers is correct.

Table 1. *Cont.*

Question Number	Questions
9	The fibers of the hypothalamic-pituitary tract originate from: a: the medial geniculate bodies; b: the mammillary bodies; c: the tuber cinereum and the mammillary bodies; d: the tuber cinereum and the supraoptic and paraventricular nuclei; e: all of the above statements are correct
10	Of the following fiber systems, only one runs through the genu of the internal capsule. Which one? a: the fibers of the corpus callosum; b: the fibers of the inferior thalamic peduncle; c: the corticospinal fibers; d: the fronto-pontine bundle; e: The corticobulbar tract
11	The "basal ganglia" encompass a group of . . .. that play a pivotal role in . . .., as well as other functions including motor learning, executive functions, and . . ..
12	The limbic lobe is a ring of ..... on the medial aspect of each hemisphere that surrounds the ..... The limbic lobe is composed of the ....., ..., and ..... gyri. The larger limbic system is involved in ..... and ..... expression.
13	....., ..., ..., ..., ... are the typical cells in the cerebrum cortex.
14	The spinocerebellar tracts convey information from ....., ..., and ..... receptors.
15	Lesions affecting the second motor neuron ..... the signal communication between the central nervous system and the muscle, resulting in ..... of the affected muscle.
16	The cochlear nuclei contain . . .. of the vestibulocochlear nerve, and the ventral cochlear nucleus in humans constitutes the main relay nucleus between first-order and second-order neurons of the . . ..
17	The . . .. part of the red nucleus is more important in humans and, like the . . .. part, is involved—though in a different way—in the . . .. exerted by the cerebellum.
18	The . . .. can be bilaterally affected in degenerative diseases such as . . .. or in . . ..; this results in difficulty swallowing due to muscle paralysis, and consequently ingested material risks entering the trachea.
19	The . . .. of the thalamus receive afferents from the . . .. of the hypothalamus and project to the cerebral cortex of the cingulate gyrus, forming part of a complex loop that also involves limbic lobe structures, known as the . . ..
20	The . . .. occupies Brodmann's area 17. It is located on the lips of the . . .. and is characterized by a very . . .. layer IV, crossed by a stripe of myelinated fibers, the stripe of the . . .. or external Baillargerstripe or . . .., which runs parallel to the cortical surface and therefore defines the striate cortex.
21	Describe the spino-bulbo-thalamo-cortical pathway and draw a diagram
22	Describe the origin and course of the median nerve. Draw a diagram.
23	Describe the morphology of the cerebellum. Then draw a diagram.
24	Which cutaneous area is innervated by the lateral femoral cutaneous nerve?
25	Describe the morphology of the internal ear. Then draw a diagram.
26	Describe the extrapyramidal system.
27	Give a description of the cortical regions of Broca's area.
28	Describe and draw a diagram illustrated the cerebellum cortex.
29	Describe the differences between neocortex and paleocortex.
30	Describe the ultrastructure of the retina.

2.3. Response Collection

Each question was submitted to all four chatbots in their open access versions. Responses were recorded verbatim without any modification.

2.4. Evaluation Process

To ensure objectivity, three independent raters with expertise in neuroanatomy conducted a blind evaluation of all responses. The evaluation focused on: accuracy: Correctness of the information provided; comprehensiveness: Depth and completeness of the explanation; relevance: Alignment of the response with the question asked; clarity: Ease of understanding and logical structure. Although the questions were collaboratively constructed by the research team, response evaluation was performed blindly, with raters unaware of chatbot identity and without access to predefined "expected" answers. Raters independently assessed responses solely on anatomical correctness and educational clarity. We acknowledge that full separation between question construction and evaluation by external assessors would further strengthen methodological rigor and should be adopted in future studies.

2.5. Scoring System: 5-Point Likert Scale

Responses were rated using a 5-point Likert scale (see Table 2). Each response was evaluated on a 5-point Likert scale, ranging from "Extremely High" to "Very Poor," based on criteria such as comprehensiveness, accuracy, and relevance (Table 2). Inter-rater reliability was deemed acceptable when the scores among the raters differed by no more than one point on the Likert scale. On the other side, a difference of 2 or more points was considered as unreliable (an event that has never occurred). To minimize potential bias, a panel of three independent experts conducted a blind assessment of each response obtained from the application. The 5-point Likert scale was intentionally applied as a global composite score, reflecting overall pedagogical usefulness rather than isolated dimensions. This mirrors real-world educational judgment, where clarity, accuracy, relevance, and completeness are often evaluated holistically. Finally, we acknowledge that the use of a single composite Likert score does not allow for the disentangling of individual dimensions such as accuracy, clarity, relevance, and comprehensiveness. Consequently, a response that is clear but partially inaccurate, or accurate but poorly structured, may receive a similar global score. While this approach was chosen to reflect overall pedagogical usefulness, future studies should employ multi-dimensional scoring systems to capture these components separately.

Table 2. 5-point Likert Scale.

Likert Point	Quality	Description
5	Extremely High quality	Totally correct, comprehensive description.
4	High quality	Mostly correct, not comprehensive description, no mistakes.
3	Moderate quality	Mostly correct, not comprehensive description, a few or minor mistakes.
2	Poor quality	Mostly wrong, poor description, many or significant mistakes.
1	Very poor quality	Totally wrong.

2.6. Statistical Analysis

Median values and interquartile ranges (IQR) were used to summarize Likert-scale data, consistent with their ordinal nature and the use of non-parametric statistical testing. Given the limited number of questions per category, statistical analyses should be interpreted as exploratory, and effect direction and consistency were prioritized over formal significance testing. Data analysis was performed using GraphPad Prism version 10 (GraphPad Software, San Diego, CA, USA). Non-parametric tests (Kruskal–Wallis with Dunn's post hoc correction) were applied to compare median Likert scores between chatbots, as these tests are appropriate for ordinal data without assuming normal distribution.

3. Results

A total of 30 neuroanatomy questions were evaluated across four categories: multiple-choice, fill-in-the-blank, descriptive, and illustration-based prompts embedded within descriptive questions. Responses from ChatGPT-4, Google Gemini, Microsoft Copilot, and Perplexity.ai were assessed by three independent raters using a 5-point Likert scale (1 = Very Poor Quality, 5 = Extremely High Quality). Evaluation criteria included accuracy, comprehensiveness, relevance, and clarity.

Overall, chatbots demonstrated a fair knowledge of human neuroanatomy. However, differences in response capabilities were highlighted not only between the different chatbots used but, even more interestingly, in relation to the type of question being asked (Supplementary Tables S1–S4).

The overall performance of the four chatbots, based on the Likert scores for all 30 questions, demonstrated a high general accuracy across all models (Table 3 and Figure 1A). In detail, ChatGPT-4 achieved the highest mean score (4.7, IQR: 4.3–5.0), closely followed by Perplexity.ai (4.3, IQR: 4.7–5.0). Microsoft Copilot (4.3, IQR: 4.5–5.0) and Google Gemini (4.3, IQR: 3.7–5.0) showed slightly lower scores. The only statistically significant differences were among ChatGPT-4 vs. Google Gemini ($p = 0.02$) and vs. Perplexity.ai ($p = 0.04$). All other comparisons were not statistically significant ($p > 0.5$).

Table 3. Evaluation of different Chatbots performances using a 5-point Likert scale.

Question	ChatGPT-4			Google Gemini			Microsoft Copilot			Perplexity.ai		
	R1	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
1	5	5	5	5	5	5	5	5	5	4	5	5
2	4	5	5	4	5	5	4	5	5	5	5	5
3	5	5	5	4	5	5	5	5	5	5	5	5
4	5	5	5	4	5	4	4	5	4	4	5	4
5	5	5	5	5	5	5	5	5	5	5	5	5
6	5	5	5	4	5	4	4	5	4	5	5	5
7	4	5	5	5	5	5	4	5	4	5	5	5
8	5	5	5	5	5	5	5	5	5	5	5	5
9	5	5	5	5	5	5	4	5	4	5	5	5
10	4	5	5	4	5	5	4	5	5	4	5	5
11	4	5	5	4	5	5	5	5	5	4	5	5
12	4	5	5	4	5	4	4	5	5	4	5	4
13	4	5	4	3	4	4	4	3	4	3	4	4

Table 3. *Cont.*

Question	ChatGPT-4			Google Gemini			Microsoft Copilot			Perplexity.ai		
	R1	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
14	4	5	5	4	3	4	4	5	4	4	5	5
15	4	5	4	4	5	4	4	5	5	4	4	4
16	3	3	3	3	2	2	5	5	5	4	5	4
17	5	5	5	5	5	5	5	5	5	5	5	5
18	5	5	5	3	3	3	5	5	5	4	3	3
19	5	5	5	5	5	5	5	5	5	4	4	4
20	4	4	4	1	1	1	5	5	5	4	4	3
21	4	4	4	5	5	5	4	4	3	4	3	3
22	4	3	3	5	5	5	4	4	4	4	5	4
23	5	5	5	4	4	4	4	4	3	5	5	5
24	5	5	4	3	2	2	4	3	3	5	5	5
25	4	5	5	3	3	3	5	5	4	4	4	4
26	4	5	4	4	4	4	4	3	3	5	5	5
27	5	5	5	4	4	4	4	4	4	5	5	5
28	4	4	4	4	4	4	4	4	4	4	5	4
29	4	4	4	4	3	3	4	4	4	5	5	5
30	5	4	4	5	5	5	4	4	4	4	5	5

Indeed, the responses to the initial 10 "multiple-choice questions" received positive evaluations from the three examiners (Table 3 and Figure 1B). The median score (and IQR) was 5 (4.7–5.0) for ChatGPT, 5.0 (4.7–5.0) for Perplexity.ai, 4.8 (4.6–5.0) for Google Gemini and 4.7 (4.3–5.0) for Microsoft Copilot. Therefore, while ChatGPT and Perplexity.AI exhibited a marginally superior performance compared to Google Gemini and Microsoft Copilot, the overall efficacy of the analysed chatbots was notably high, demonstrating a robust capability to provide accurate and relevant information (ChatGPT-4 vs. Microsoft Copilot $p = 0.04$, for the other comparisons $p > 0.05$). The difference in evaluation is related to the completeness of the information from some chatbots that didn't just give the correct answer but went further to provide a neuroanatomical description.

The evaluation of questions 11 through 20, which were of the "fill-in-the-blank" format, revealed a similar performance pattern (Table 3 and Figure 1C). Specifically, ChatGPT and Microsoft Copilot emerged as the most proficient, achieving a median score of 4.7 (IQR: 4.3–5.0) and 5 (IQR: 4.6–5.0), respectively. Conversely, Perplexity.ai (4.2, IQR: 3.7–4.7) and Google Gemini (3.8, IQR: 2.3–4.4) demonstrated a slightly comparatively reduced level of precision. In particular, the worst performances were of Google Gemini ($p = 0.03$, and $p = 0.02$ respectively vs. ChatGPT-4 and Microsoft Copilot. The differences were not statistically significant ($p > 0.05$). These findings underscore the nuanced variability in chatbot performance across different question formats, highlighting the importance of evaluating AI capabilities within specific task domains.

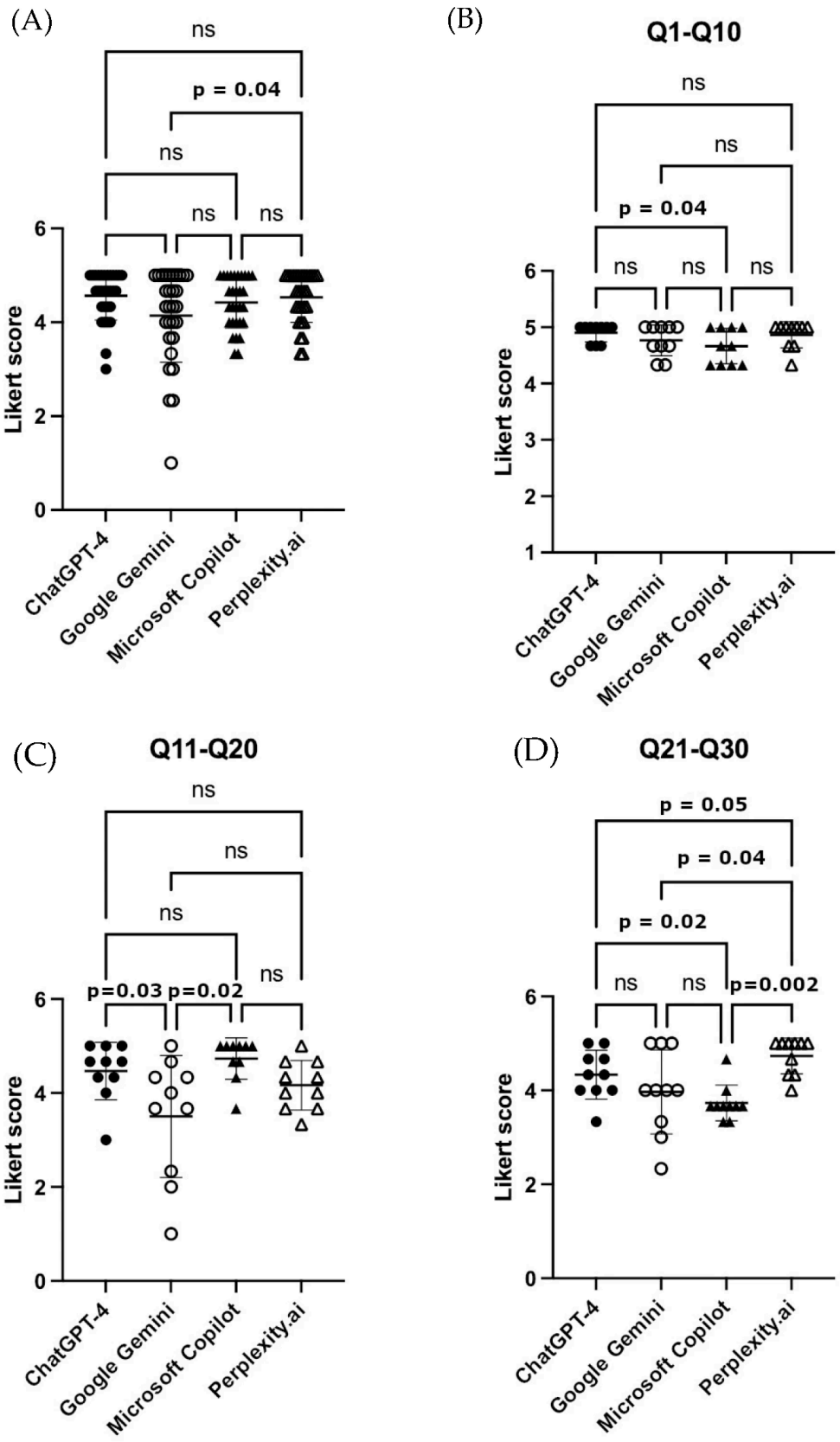


Figure 1. Overall Chatbot Performance (Mean Likert Score). (**A**) Chart comparing the mean Likert scores of four AI chatbots, ChatGPT-4, Perplexity.ai, Microsoft Copilot, and Google Gemini, based on their responses to 30 neuroanatomy questions. ChatGPT-4 and Perplexity.ai achieved the highest mean scores, followed by Microsoft Copilot and Google Gemini. (**B**) Chart comparing the mean Likert

scores of four AI chatbots, ChatGPT-4, Perplexity.ai, Microsoft Copilot, and Google Gemini, based on their responses to Q1–Q10 neuroanatomy questions. (**C**) Chart comparing the mean Likert scores of four AI chatbots, ChatGPT-4, Perplexity.ai, Microsoft Copilot, and Google Gemini, based on their responses to Q11–Q20 neuroanatomy questions. (**D**) Chart comparing the mean Likert scores of four AI chatbots, ChatGPT-4, Perplexity.ai, Microsoft Copilot, and Google Gemini, based on their responses to Q21–Q30 neuroanatomy questions. Error bars represent standard deviation, indicating variability in performance across question types. Numbers indicate p values (significant $p < 0.05$).

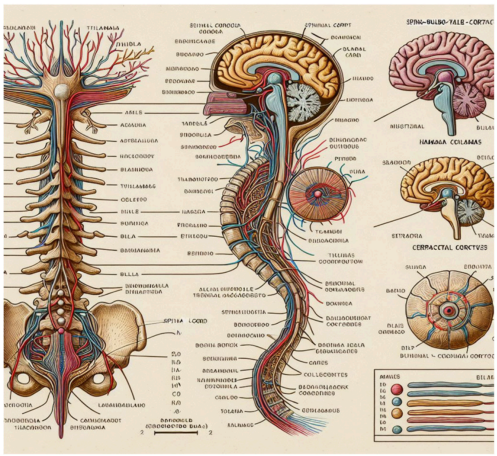
The comparison of the performances of the different bots in answering "describe and explain"-type questions (from 21 to 30) was notably poorer (Tables S1–S3 and summarised in Table 3). This part of the study was certainly the one that put the different bots under the most stress, especially for the part concerning the "Illustrate: generate or suggest appropriate diagrams and visual aids". The results, therefore, highlighted a differential capacity among the chatbots to respond to this type of question. Perplexity.ai proved to be the best in formulating accurate neuroanatomical descriptions (score 5, IQR: 4.3–5.0), followed by ChatGPT-4 (4.3, IQR: 4.0–4.8). On the contrary, Google Gemini (score 4.0, IQR: 3.3–5.0) and Microsoft Copilot (3.6, IQR: 3.6–3.8) provided answers that were not always consistent with the question. In certain instances, the information provided by the chatbots was suboptimal. Thus, the best performer was Perplexity.ai ($p = 0.05$, $p = 0.04$ and $p = 0.002$ respectively vs. ChatGPT-4, Google Gemini, and Microsoft Copilot). Finally, the other differences were not statistically significant ($p > 0.05$).

It is widely recognized that the study of anatomy cannot be done merely from a text. Images are essential to fully understand the course of a nerve, the relationships between various structures, and so on. Thus, students could read a description, but the easier way to understand is to see a picture.

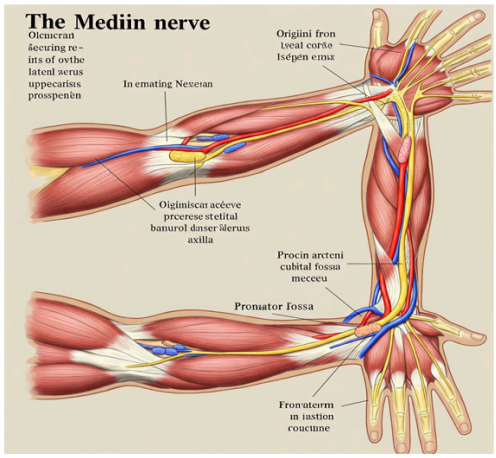
When queried about human anatomy, the chatbots provided textual information of varying quality. However, a significant limitation was the inability of the majority to generate visual representations. The absence of visual representations poses a significant challenge because the study of human anatomy fundamentally relies on the precise understanding of the morphology and spatial relationships of internal structures. In simpler terms, to truly grasp how our bodies work, we need to know not only what things are, but also where they are and what shape they have. That's why images are so fundamental.

The free version of ChatGPT-4 and Google Gemini are unable to draw any type of image. Perplexity.ai, although unable to draw, often inserts links to images on the web. Finally, Microsoft Copilot proposes images that are often fanciful, with incorrect anatomical details, which can be misleading, especially in educational or medical contexts. Even the captions of the proposed images are characterized by errors. Fortunately, these errors are so evident that even a beginner student in anatomy can notice them (Figure 2).

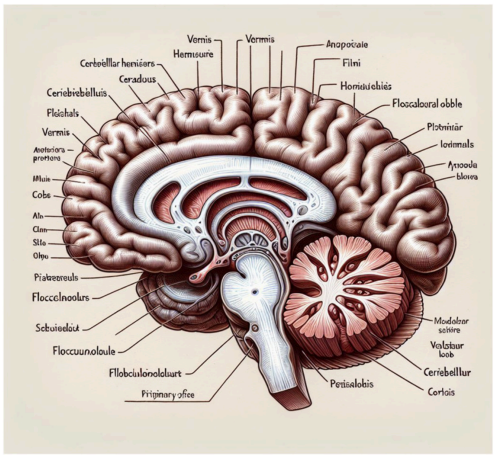
All these tests were performed during the last April 2025. We recently performed (December 2025) a new evaluation on Microsoft Copilot capabilities to draw anatomical images. The aim was to verify a possible improvement of this Chatbot. As shown in Figure 3, Microsoft Copilot's ability to draw neuroanatomical images remains very imprecise and often "imaginative." Even after more than six months, its ability to represent anatomical images has not improved, highlighting a significant weakness in this system. Fortunately, the images are so "imaginative" and clearly incorrect that they could not lead a student to interpret them as accurate representations of reality.



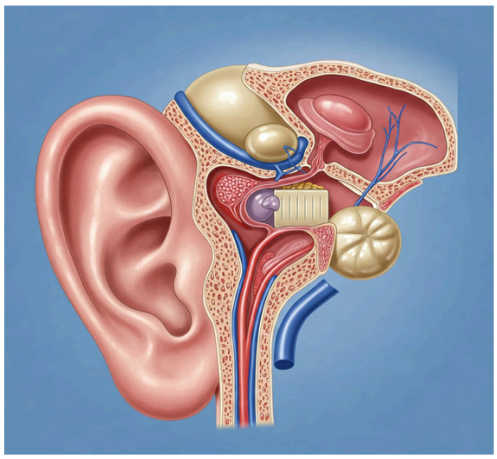
(**A**)



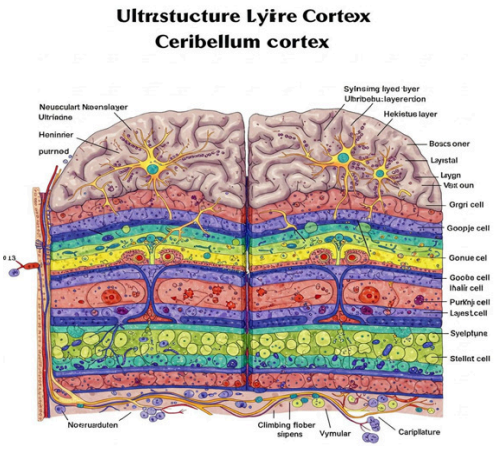
(**B**)



(**C**)



(**D**)



(**E**)

Figure 2. A graphical attempt to represent a neuroanatomical structure (April 2025). Microsoft Copilot tried to draw the spino-bulbo-thalamo-cortical pathway (**A**), the origin and the course of the median nerve (**B**), the morphology of the cerebellum (**C**), the inner ear (**D**), and the microscopic morphology of the cerebellum's cortex (**E**). It's clear that the pictures don't make sense.

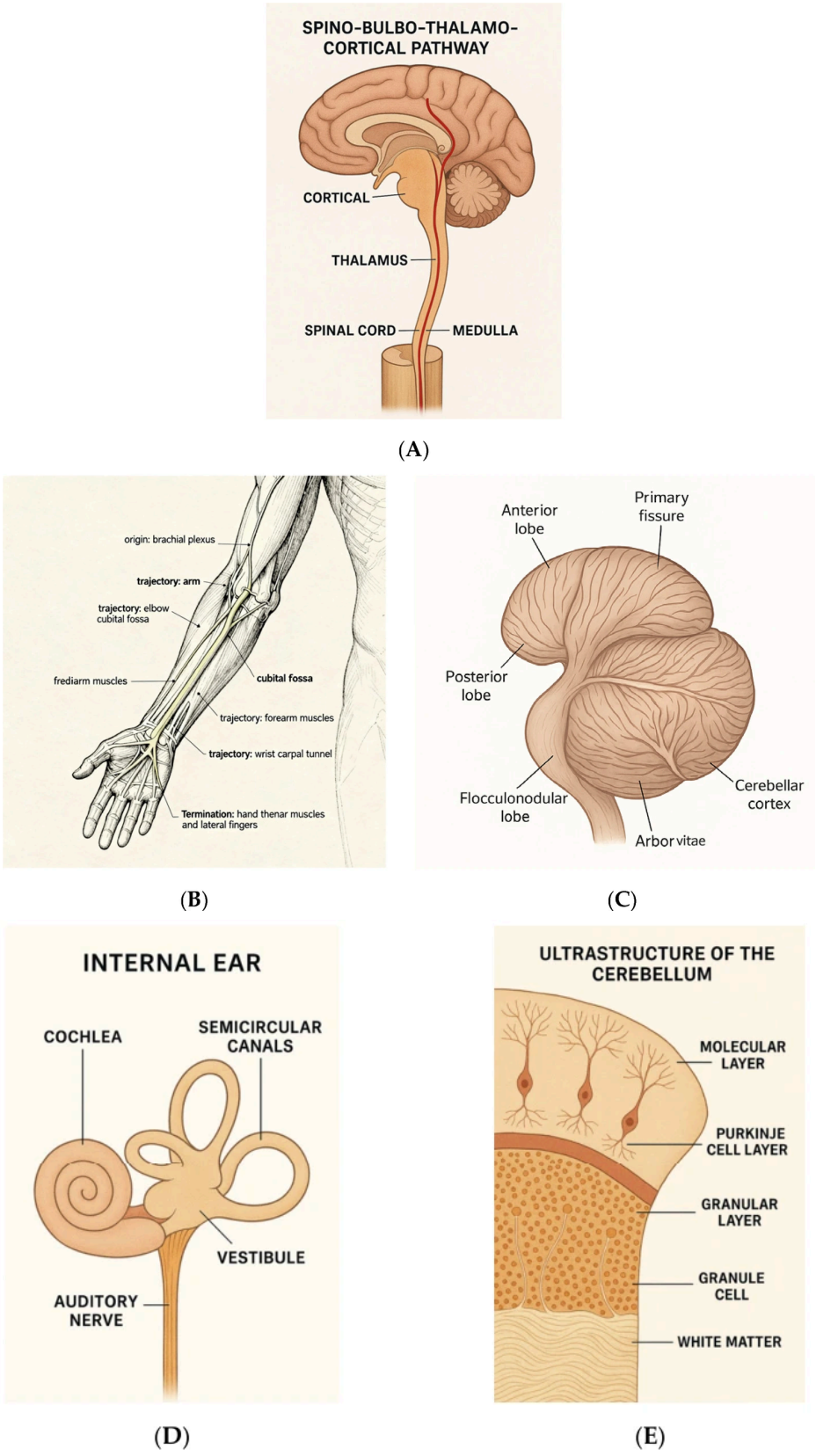


Figure 3. A graphical attempt to represent a neuroanatomical structure (December 2025). Microsoft Copilot tried to draw the spino-bulbo-thalamo-cortical pathway (**A**), the origin and the course of the median nerve (**B**), the morphology of the cerebellum (**C**), the inner ear (**D**), and the microscopic morphology of the cerebellum's cortex (**E**). Although the graphic results have improved, the output is still far from being acceptable and instructive for a medical student.

4. Discussion

Since its introduction at the end of 2022, ChatGPT-4 marked a significant milestone in the application of AI within higher education. Owing to its advanced natural language processing capabilities, the model rapidly gained widespread popularity among students worldwide. However, the growing adoption of ChatGPT-4 by the student population has sparked considerable global debate, eliciting both enthusiasm and scepticism [22]. These dynamics have prompted a systematic investigation aimed at assessing the effectiveness and accuracy of the system's responses to potential neuroanatomy-related questions.

The present study evaluated the performance of four widely used AI chatbots, ChatGPT-4, Google Gemini, Microsoft Copilot, and Perplexity.ai, in answering neuroanatomy questions of varying complexity. In this study, Likert scores were intended to capture expert-perceived response quality and pedagogical plausibility rather than educational effectiveness. While such evaluations may inform the potential suitability of chatbot outputs as learning materials, they should not be interpreted as evidence of learning gains, knowledge retention, or clinical competence.

4.1. Comparison with Previous Literature

Our findings align with earlier reports emphasizing the growing role of AI-based large language models (LLMs) in medical education, particularly in anatomy, where rapid access to structured information can enhance self-directed learning and examination preparation [3,10]. Similar to the results of Singal et al. [17], who reported high accuracy rates for AI platforms in anatomy MCQs, our study confirms that LLMs excel in structured question formats. However, consistent with observations by Al-Khater [4] and Bolgova et al. [8], the inability of these models to accurately interpret or generate anatomical illustrations remains a critical shortcoming.

Finally, our results are consistent with those reported by Singal and Goyal [17], who observed near-perfect accuracy among LLMs when solving anatomy MCQs, with Google Gemini models performing slightly better than ChatGPT-4 and DeepSeek. AI chatbots demonstrated strong performance in text-based neuroanatomy questions but persistent limitations in spatial and visual representation. Based on these findings, such tools may serve as supplementary resources to reinforce theoretical knowledge, while traditional instruction and validated atlases remain essential, particularly for spatially complex anatomy.

4.2. Performance Trends and Reliability

ChatGPT-4 and Perplexity.ai achieved the highest overall Likert scores, indicating strong reliability for text-based queries. Microsoft Copilot and Google Gemini followed closely, though Gemini exhibited greater variability, suggesting inconsistent performance across question types. These findings echo prior studies where ChatGPT-4 demonstrated superior consistency compared to other platforms in anatomy-related tasks [5,10]. Inter-rater agreement in our study was high, reinforcing the robustness of the evaluation process.

4.3. Challenges in Visual and Applied Anatomy

Illustration-based prompts received the lowest ratings, underscoring a persistent gap in AI's ability to handle spatially complex anatomical information. This limitation has been highlighted in previous research, where LLMs struggled with radiologic anatomy and surface marking tasks [4,10]. Although Microsoft Copilot attempted image generation, the outputs were anatomically inaccurate, raising concerns about potential misinformation if such tools are used without expert oversight.

Of note, we observed the lack of improvement and/or learning of Chatbots over the past six months in the figurative representation abilities of neuroanatomy. This implies

that students should carefully evaluate images created by Chatbots. On the other hand, the errors are so manifest that even a beginner student would hardly be misled. In conclusion, we believe that, for those approaching the study of neuroanatomy for the first time, using an atlas or textbook remains the best solution. Paradoxically, Google search and chatbots that retrieve images from the internet or from scientific articles on PubMed are more reliable. Of note, the observation regarding lack of improvement in figurative representations over six months is based on informal longitudinal observation rather than a controlled comparative analysis and should be interpreted accordingly.

4.4. Educational Implications

The high performance of chatbots in factual domains suggests their potential as supplementary tools for reinforcing theoretical knowledge. Their speed and accessibility can support active learning strategies, particularly for first-year medical students who often face cognitive overload in anatomy [2–4,9–13,16,17]. However, reliance on these tools without critical appraisal may propagate errors, especially in clinically relevant or visually dependent topics. Therefore, integration of AI into anatomy curricula should be accompanied by structured guidance and expert validation [2–4,9–13,16,17].

5. Conclusions

The present study highlights the growing potential of AI-based chatbots as supplementary tools in neuroanatomy education. Among the evaluated platforms, ChatGPT-4 and Perplexity.ai demonstrated the highest overall performance, achieving strong Likert ratings for accuracy, clarity, and relevance in text-based question formats. Microsoft Copilot and Google Gemini also performed well, though with slightly greater variability. These findings reinforce the capability of LLMs to support factual learning and rapid information retrieval, which can be particularly beneficial for early-stage medical learners facing extensive cognitive demands.

However, the persistent limitations observed in illustration-based and higher-order descriptive tasks underscore a critical gap in current AI technology. Neuroanatomy, as a discipline, relies heavily on spatial reasoning and visual interpretation, domains where all evaluated chatbots underperformed. This shortcoming not only restricts their role as standalone educational resources but also raises concerns about potential misinformation if such tools are used without expert oversight.

From an educational perspective, AI chatbots can complement traditional teaching by providing immediate, accessible explanations and reinforcing theoretical knowledge. Yet, their integration into curricula must be approached cautiously, with structured guidance and continuous validation by subject matter experts. Future advancements should prioritize the incorporation of validated visual resources, adaptive feedback mechanisms, and domain-specific fine-tuning to bridge the current performance gap. Of note, chatbot performance scores in this study reflect response quality under expert evaluation, not direct measures of student learning, retention, or clinical application. Educational utility should therefore not be inferred directly from these scores.

In conclusion, while AI chatbots represent a promising adjunct in anatomy education, they cannot replace the irreplaceable value of hands-on dissection, interactive visualization, and expert-led instruction. Their optimal role lies in augmenting, not substituting, traditional pedagogical strategies, ensuring that technological innovation enhances, rather than compromises, the integrity of medical education.

6. Future Directions in AI-Assisted Anatomy Education

The findings of this study emphasize the need for ongoing refinement of AI-based tools in anatomy education. Future development should highlight the integration of validated, peer-reviewed visual resources to overcome current limitations in anatomical illustration. Training AI with neuroanatomy data and medical ontologies could make it more accurate and reduce errors. Reliability could be strengthened through adaptive feedback systems that allow users to flag mistakes and receive corrected explanations, promoting active learning while improving model performance. Evaluations should also expand to include problem-solving, clinical reasoning, and applied anatomy scenarios, offering a more comprehensive view of AI's educational potential. Comparative studies of free versus premium models are warranted to inform institutional adoption. Finally, ethical and pedagogical oversight, combined with larger question banks and image-based assessments, will be essential to ensure safe, effective, and equitable integration of AI into anatomy curricula.

7. Methodological Considerations and Limitations

We recognize that the same research group was involved in both question construction and response evaluation, which may raise concerns regarding circular assessment. However, the primary objective of this study was not to validate the absolute correctness of individual questions, but to perform a comparative evaluation across different chatbot systems using identical prompts. Because all chatbots were assessed on the same questions under blinded conditions, any potential bias related to question design would apply equally to all systems and is therefore unlikely to systematically favor one chatbot over another. Nonetheless, we acknowledge that future studies would benefit from complete separation between question development and evaluation, ideally involving external raters, to further strengthen methodological independence.

Inter-rater reliability was assessed descriptively by agreement within one Likert point rather than by formal coefficients such as Cohen's weighted kappa or intraclass correlation. This approach was chosen because the scoring system was designed as a global pedagogical judgment rather than as independent dimensional ratings and because the limited number of raters and ordinal nature of the scale reduced the interpretability of parametric reliability estimates. Nonetheless, we acknowledge that the absence of a formal inter-rater reliability coefficient represents a methodological limitation. Future studies should report standardized reliability indices to further strengthen reproducibility.

A further limitation of this study is the relatively small number of questions (n = 30), which results in limited statistical power, particularly when analyses are stratified by question format. Consequently, the study is underpowered to detect small effect sizes, and p-values should be interpreted with caution. For this reason, the present analyses are intended to be exploratory and comparative in nature, emphasizing consistency and direction of differences between chatbots rather than definitive statistical significance. Future studies employing larger and systematically sampled question banks will be required to enable adequately powered inferential analyses and robust effect size estimation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17050475/s1, Table S1. Full textual responses generated by ChatGPT 4 to the 30 neuroanatomy questions used in this study. Table S2. Full textual responses generated by Google Gemini to the 30 neuroanatomy questions used in this study. Table S3. Full textual responses generated by Microsoft Copilot to the 30 neuroanatomy questions used in this study. Table S4. Full textual responses generated by Perplexity.ai to the 30 neuroanatomy questions used in this study.

Author Contributions: Conceptualization, A.N., S.N. and D.S.; Methodology, S.N.; Formal analysis, S.N.; Investigation, A.N.; Data curation, A.N. and D.S.; Writing—original draft, A.N. and S.N.; Writing—review & editing, D.S.; Supervision, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahn, S. The impending impacts of large language models on medical education. *Korean J. Med. Educ.* **2023**, *35*, 103–107. [CrossRef]
2. Totlis, T.; Natsis, K.; Filos, D.; Ediaroglou, V.; Mantzou, N.; Duparc, F.; Piagkou, M. The potential role of ChatGPT and artificial intelligence in anatomy education: A conversation with ChatGPT. *Surg. Radiol. Anat.* **2023**, *45*, 1321–1329. [CrossRef] [PubMed]
3. Mogali, S.R. Initial impressions of ChatGPT for anatomy education. *Anat. Sci. Educ.* **2024**, *17*, 444–447. [CrossRef] [PubMed]
4. Mantzou, N.; Ediaroglou, V.; Drakonaki, E.; Syggelos, S.A.; Karageorgos, F.F.; Totlis, T. ChatGPT efficacy for answering musculoskeletal anatomy questions: A study evaluating quality and consistency between raters and timepoints. *Surg. Radiol. Anat.* **2024**, *46*, 1885–1890. [CrossRef] [PubMed]
5. Sallam, M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
6. Lee, H. The rise of ChatGPT: Exploring its potential in medical education. *Anat. Sci. Educ.* **2024**, *17*, 926–931. Erratum in *Anat. Sci. Educ.* **2024**, *17*, 1779. https://doi.org/10.1002/ase.2496. [CrossRef]
7. Bolgova, O.; Ganguly, P.; Mavrych, V. Comparative analysis of LLMs performance in medical embryology: A cross-platform study of ChatGPT, Claude, Gemini, and Copilot. *Anat. Sci. Educ.* **2025**, *18*, 718–726. [CrossRef]
8. Boscardin, C.K.; Gin, B.; Golde, P.B.; Hauer, K.E. ChatGPT and generative artificial intelligence for medical education: Potential impact and opportunity. *Acad. Med.* **2024**, *99*, 22–27. [CrossRef]
9. Al-Khater, K.M.K. Comparative assessment of three AI platforms in answering USMLE step 1 anatomy questions or identifying anatomical structures on radiographs. *Clin. Anat.* **2025**, *38*, 186–199. [CrossRef]
10. Mavrych, V.; Ganguly, P.; Bolgova, O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: Comparative analysis. *Clin. Anat.* **2025**, *38*, 200–210. [CrossRef] [PubMed]
11. Manavalan, M.S.; Sundaramurthi, I.; Surapaneni, K.M. Assessment of the performance of ChatGPT in human anatomy of first professional MBBS Competency Based Medical Education (CBME) undergraduate. *Indian J. Anat.* **2024**, *13*, 17–26. [CrossRef]
12. Saluja, S.; Tigga, S.R. Capabilities and Limitations of ChatGPT in Anatomy Education: An Interaction with ChatGPT. *Cureus* **2024**, *16*, e69000. [CrossRef] [PubMed]
13. Abdellatif, H.; Al Mushaiqri, M.; Albalushi, H.; Al-Zaabi, A.A.; Roychoudhury, S.; Das, S. Teaching, Learning and Assessing Anatomy with Artificial Intelligence: The Road to a Better Future. *Int. J. Environ. Res. Public Health* **2022**, *19*, 14209. [CrossRef] [PubMed]
14. Sarker, I.H. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Comput. Sci.* **2022**, *3*, 158. [CrossRef]
15. Korteling, J.E.; van de Boer-Visschedijk, G.C.; Blankendaal, R.A.; Boonekamp, R.C.; Eikelboom, A.R. Human-versus artificial intelligence. *Front. Artif. Intell.* **2021**, *4*, 622364. [CrossRef]
16. Talan, T.; Kalınkara, Y. The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course. *Uluslararası Yönetim Bilişim Sist. Ve Bilgi. Bilim. Derg.* **2023**, *7*, 33–40. [CrossRef]
17. Singal, A.; Goyal, S. Comparative evaluation of AI platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o" in solving multiple-choice questions from different subtopics of anatomy. *Surg. Radiol. Anat.* **2025**, *47*, 193–202. [CrossRef]
18. García-López, I.M.; Trujillo-Liñán, L. Ethical and regulatory challenges of Generative AI in education: A systematic review. *Front. Educ.* **2025**, *10*, 1565938. [CrossRef]
19. Hua, S.; Jin, S.; Jiang, S. The limitations and ethical considerations of ChatGPT. *Data Intell.* **2024**, *6*, 201–239. [CrossRef]
20. Abbas, A.; Rehman, M.S.; Rehman, S.S. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus* **2024**, *16*, e55991. [CrossRef] [PubMed]

21. Wang, K.D.; Burkholder, E.; Wieman, C.; Salehi, S.; Haber, N. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Front. Educ.* **2024**, *8*, 1330486. [CrossRef]
22. Aristovnik, A.; Ravšelj, D.; Keržič, D.; Tomaževič, N.; Umek, L.; Brezovar, N.; Kessler, S.H.; Schäfer, M.; Fišer, Z. Higher Education Students' Evolving Perceptions of ChatGPT: Global Survey Data from the Academic Year 2024–2025. *Mendeley Data V1* **2025**, *1*. [CrossRef]