VETERİNER HEKİMLER DERNEĞİ DERGİSİ
JOURNAL OF THE TURKISH VETERINARY MEDICAL SOCIETY
ISSN: 0377-6395
e-ISSN: 2651-4214
Dergi ana sayfası /*Journal homepage:* http://dergipark.org.tr/vetheder
OPEN ACCESS

10.33188/vetheder.1805359

Araştırma Makalesi / Research Article

A comparative evaluation of AI chatbots in veterinary anatomy: Performance of ChatGPT, Gemini and DeepSeek models

Ezgi Deniz MAVİLİ [1,6,a], ***Barış BATUR*** [1,6,b], ***Aytaç AKÇAY*** [7,c], ***Çağdaş OTO*** [1,2,3,4,5,d]

[1] *Ankara University Faculty of Veterinary Medicine, Department of Anatomy, Ankara, Turkiye*
[2]*Ankara University Medical Design Research and Application Center MEDITAM, Ankara, Turkiye*
[3]*Integrated Technologies Research Center (BÜTAM), Ankara University, Ankara, Turkiye.*
[4]*Ankara University Interventional MRI Clinical R&D Institute, Ankara, Turkiye*
[5]*Ankara University Biotechnology Institute, Ankara, Turkiye.*
[6]*Graduate School of Health Sciences, Ankara University, Ankara, Turkiye.*
[7] *Ankara University Faculty of Veterinary Medicine, Department of Biostatistic, Ankara, Turkiye.*
0009-0002-3565-8825 [a]; *0000-0001-9669-9917*[b]; *0000-0001-6263-5181*[c] ; *0000-0002-2727-3768* [d]

MAKALE BİLGİSİ / ARTICLE INFORMATION:

Geliş / Received:
16 Ekim 25
16 October 25

Revizyon/Revised:
18 Aralık 25
18 December 25

Kabul / Accepted:
30 Aralık 25
30 December 25

Anahtar Sözcükler:
Yapay
Zeka
Büyük Dil Modelleri

Keywords:
Artificial
Intelligence
Large Language Models

©2026 The Authors. Published by Veteriner Hekimler Derneği. This is an open access article under CC-BY-NC license. (https://creativecommons.org/licenses/by-nc/4.0)

ABSTRACT:

This study aimed to evaluate the reliability and accuracy of four AI chatbots—ChatGPT-3.5, ChatGPT-4.0, Gemini 2.5 Flash, and DeepSeek-V3—in the field of veterinary anatomy. A total of 85 multiple-choice questions encompassing major anatomical systems were presented individually to each model under identical conditions. Responses were evaluated for accuracy, and success rates were calculated as percentages. Statistical differences among models were analyzed using the Pearson chi-square test ($p<0.05$). The results indicated that Gemini 2.5 Flash achieved the highest accuracy rate (85.88%), followed by ChatGPT-4.0 (85.53%), DeepSeek-V3 (84.71%), and ChatGPT-3.5 (82.35%). Despite these variations, the differences were not statistically significant ($\chi^2=0.629$, $p=0.890$). Qualitative analysis revealed differences in explanatory depth: ChatGPT-4.0 and Gemini 2.5 Flash provided corrective feedback for incorrect options, while DeepSeek-V3 and ChatGPT-3.5 focused mainly on correct answers. Gemini 2.5 Flash additionally incorporated visual aids, though some were based on human rather than veterinary anatomy. Overall, while all evaluated AI chatbots demonstrated a substantial capacity for accurate anatomical reasoning, their explanatory styles and supporting materials varied.

Veteriner anatomisinde yapay zeka sohbet robotlarının karşılaştırmalı değerlendirmesi: ChatGPT, Gemini ve DeepSeek modellerinin performansı

ÖZET:

Bu çalışma, veteriner anatomisi alanında dört yapay zeka sohbet robotunun (ChatGPT-3.5, ChatGPT-4.0, Gemini 2.5 Flash ve DeepSeek-V3) güvenilirliğini ve doğruluğunu değerlendirmek amacıyla yapılmıştır. Başlıca anatomik sistemleri kapsayan toplam 85 çoktan seçmeli soru, aynı koşullar altında her modele ayrı ayrı sunulmuştur. Yanıtlar doğruluk açısından değerlendirilmiş ve başarı oranları yüzde olarak hesaplanmıştır. Modeller arasındaki istatistiksel farklılıklar Pearson ki-kare testi ($p<0,05$) kullanılarak analiz edilmiştir. Sonuçlar, Gemini 2.5 Flash'ın en yüksek doğruluk oranını (%85,88) elde ettiğini, onu ChatGPT-4.0 (%85,53), DeepSeek-V3 (%84,71) ve ChatGPT-3.5 (%82,35) izlediğini gösterdi. Bu farklılıklara rağmen, farklar istatistiksel olarak anlamlı değildi ($\chi^2=0,629$, $p=0,890$). Niteliksel analiz, açıklayıcı derinlik açısından farklılıklar ortaya koydu: ChatGPT-4.0 ve Gemini 2.5 Flash, yanlış seçenekler için düzeltici geri bildirim sağlarken, DeepSeek-V3 ve ChatGPT-3.5 esas olarak doğru cevaplara odaklandı. Gemini 2.5 Flash ayrıca görsel yardımcılar da kullanmıştır, ancak bunların bazıları veteriner anatomisi yerine insan anatomisine dayanmaktadır. Genel olarak, değerlendirilen tüm AI sohbet robotları doğru anatomik muhakeme konusunda önemli bir kapasite sergilemiş olsa da, açıklama stilleri ve destekleyici materyalleri farklılık göstermektedir.

How to cite this article: Mavili ED, Batur B, Akçay A, Oto Ç. A comparative evaluation of AI chatbots in veterinary anatomy: Performance of ChatGPT, Gemini and DeepSeek models. Vet Hekim Der Derg. 2026; 97(1):47-51. Doi: 10.33188/vetheder.1805359

* *Sorumlu Yazar e-posta adresi / Corresponding Author e-mail address: ezgidenizmavili@gmail.com*

1. INTRODUCTION

In recent years, significant progress in educational methodologies has profoundly reshaped the teaching of anatomy, with the integration of advanced digital technologies, interactive e-learning platforms, and a wide array of online resources complementing traditional didactic approaches to create a more engaging, flexible, and student-centered learning environment (1). The integration of artificial intelligence into anatomy education aids problem-solving and enables the execution of difficult tasks with greater precision and adaptability, eventually altering traditional learning techniques and enhancing educational outcomes (2). GPT-based conversational agents have the potential to facilitate dynamic, interactive dialogue with students, enabling them to inquire about specific anatomical structures or systems while offering comprehensive explanations, precise definitions, and in-depth descriptions that support a deeper and more accessible understanding of the complexities of veterinary anatomy through an engaging conversational format (3). The growing prominence of ChatGPT and other AI tools powered by large language models has intensified scholarly discourse on issues of reliability, validity, and ethics, particularly due to their occasional production of inaccurate content, omission of essential information, and generation of questionable or non-existent references (4). In light of the escalating discourse surrounding the subject, this study has been conducted with the objective of ascertaining the reliability of artificial intelligence in the domain of veterinary anatomy. One study conducted outside the field of veterinary anatomy has revealed that both versions 3.5 and 4.0 of ChatGPT do not provide sufficiently accurate or reliable anatomical information regarding the scalenovertebral triangle and therefore cannot be considered a reliable reference for this specific anatomical topic (5). Concerns have emerged that the well-documented limitations of ChatGPT-based outputs may negatively influence students understanding of anatomy and their ability to apply anatomical knowledge in clinical contexts (4). To test the accuracy performance of artificial intelligence chatbots in the field of clinical anatomy, the Gemini, Claude, and ChatGPT 3.5 models were subjected to USMLE Step 1 anatomy questions (2). Although various performance tests for artificial intelligence models already exist in the anatomy literature, more are needed in comprehensive fields such as veterinary anatomy. This study comparatively evaluated the veterinary anatomy knowledge of four AI chatbots ChatGPT-3.5, ChatGPT-4.0, Gemini 2.5 Flash, and DeepSeek-V3 using multiple-choice questions.

2. MATERIAL AND METHODS

A total of 85 multiple-choice questions were prepared, each consisting of five options with only one correct answer. The 85 multiple-choice questions were designed as follows: 9 questions pertain to osteology, 9 to the muscular system, 3 to the joint system, 9 to the respiratory system, 10 to the digestive system, 9 to the urogenital system, 9 to the circulatory system, 9 to the nervous system, 9 to the sensory organs, and 9 to avian anatomy. These questions were developed by faculty members who have over 15 years of experience teaching veterinary anatomy at veterinary faculty. The questions covered various anatomical systems relevant to veterinary education, including the skeletal system, joints, muscles, urogenital system, respiratory system, digestive system, circulatory system, and nervous system. All questions were presented to each chatbot individually, without any additional prompts or contextual information. The intention was to assess the models based solely on their existing knowledge and reasoning capabilities. Each chatbot was asked the same set of questions under identical conditions to ensure consistency in the evaluation process. Once all responses were collected, each answer was checked against the predetermined correct option. Answers were marked as either correct or incorrect; no partial credit or interpretive assessment was applied. The evaluation focused strictly on accuracy. To determine the success rate of each chatbot, the number of correct answers was divided by the total number of questions (85), and the result was expressed as a percentage. No external sources, tools, or human corrections were used during the process. The differences in the correct response rates of the artificial intelligence programs were tested for statistical significance using the Pearson chi-square test. A significance level of $p<0.05$ was accepted. Statistical analyses were performed using the SPSS 30.0 software package.

3. RESULTS

The evaluation of the selected AI chatbots revealed notable differences in their performance when tested with a standardized set of veterinary anatomy multiple-choice questions. Among the evaluated models, Gemini 2.5 Flash demonstrated the highest level of accuracy, achieving a success rate of 85.88%. Closely following Gemini 2.5 Flash, ChatGPT 4.0 reached a success rate of 85.53%, showing a very narrow margin between the top two models. DeepSeek V3 achieved a success rate of 84.71%, positioning it third among the models tested. In contrast, ChatGPT 3.5 produced the lowest success rate, with an accuracy of 82.35%.

The comparative analysis of the explanatory styles of the four evaluated models demonstrated clear differences in the way information was presented (Table 1). All chatbots consistently provided explanations related to the correct answer option, confirming their baseline capacity to identify and justify the correct response. However, variations were observed in the treatment of incorrect alternatives. While both ChatGPT 4.0 and Gemini 2.5 Flash offered corrective feedback by addressing why the other options were unsuitable, DeepSeek V3 and ChatGPT 3.5 limited their outputs to explanations focused exclusively on the correct choice. Notably, Gemini 2.5 Flash distinguished itself from the other models by incorporating visual material in its responses. Although the question asked pertained to veterinary anatomy, visual support related to human anatomy was provided.

In terms of correct and incorrect response rates, the differences among the artificial intelligence programs were not found to be statistically significant (χ^2=0.629, p=0.890).

Table 1: Comparative explanatory features of the evaluated AI chatbots.

Tablo 1: *Değerlendirilen AI sohbet robotlarının karşılaştırmalı açıklayıcı özellikleri.*

Model	Information about the correct option	Information about the wrong options	Visual support for the correct answer
ChatGPT 3.5	✓	✗	✗
ChatGPT 4.0	✓	✓	✗
DeepSeek V3	✓	✗	✗
Gemini Flash 2.5	✓	✓	✓

4. DISCUSSION AND CONCLUSION

This study was designed to address a central research question: To what extent can contemporary artificial intelligence chatbots provide accurate and reliable knowledge in the specialized field of veterinary anatomy? While existing literature has evaluated large language models in relation to general medical education and human anatomy, investigations focusing specifically on veterinary anatomy remain limited. By comparing the performance of four AI models—ChatGPT 3.5, ChatGPT 4.0, Gemini 2.5 Flash, and DeepSeek V3—on a standardized set of multiple-choice questions developed by veterinary anatomists, this study sought to determine their relative accuracy, explanatory quality, and potential usefulness as supportive tools in anatomy education.

The LLM models examined in this study exhibited significant variation in their responses to veterinary anatomy questions, particularly regarding the comprehensiveness and clarity of the explanations provided.

For ChatGPT-4, one of the models tested, the published technical report has indicated that it achieved high success rates in examinations conducted in the United States. These examinations include the multistate bar exam, the academic proficiency test, and the graduate school entrance examination (6). Given that GPT-4 has achieved a certain level of success in these examinations, in our study was designed with an approach aimed at evaluating the success rates of multiple-choice questions that could potentially be used to train students. In this way, the reliability of artificial intelligence chatbots in delivering conceptual knowledge within the domain of veterinary anatomy could be

systematically evaluated. Unlike the human mind, LLMs lack the ability to evaluate evidence or reason (7), and therefore it is thought that they could be used as an auxiliary tool in the field of veterinary anatomy.

The success rates and response characteristics we obtained from four different models can be used as a helpful resource for both students and instructors when selecting an auxiliary tool model in training and counselling. These models have been selected because they enable effective information extraction and advanced analytical capabilities from large and complex data sets (8). Although these models exhibit advantageous features (9), the inherent constraints of artificial intelligence differ among models (4). A comparison between ChatGPT-4 and Gemini revealed that ChatGPT-4o Mini was more successful in analytical and application-level questions, whereas Gemini demonstrated stronger performance in descriptive and comprehension-level questions (1).In our study, a 3.18% difference in performance was observed between the ChatGPT 3.5 and ChatGPT 4.0 models, despite both belonging to the same brand. When evaluated alongside veterinary anatomy tests conducted with models from other companies such as DeepSeek and Gemini, this difference highlighted the distinct advantages and disadvantages of the models used.

Several limitations must be acknowledged in interpreting these results. First, the evaluation was confined to a fixed set of 85 multiple-choice questions prepared by veterinary anatomists. While this ensured accuracy and relevance, the findings may not fully capture the performance of the models across different question formats, such as open-ended or clinically oriented problem-solving tasks. Second, only one type of question format was employed, and no prompts or contextual cues were introduced; the performance of these models might vary under different testing conditions or with alternative assessment methods. Third, the explanatory outputs were not evaluated for pedagogical effectiveness or student comprehension, but only for their accuracy in relation to the correct option. Additionally, while statistical analysis was applied to compare overall accuracy, the qualitative aspects of the explanations—such as clarity, comprehensiveness, or potential to mislead—were not systematically assessed beyond descriptive observation. Future studies should therefore extend the evaluation to broader question types, explore user-centered testing with students, and investigate the long-term educational impact of integrating AI chatbots into veterinary anatomy curricula.

Since prompt design can influence the performance of AI chatbots in different ways (10), no additional prompts were included in our multiple-choice questions; instead, only the question stem and five answer options were presented separately to the AI chatbots. We anticipated that the performance of language models might vary when they generate and answer their own questions, compared to when they are evaluated using expert-designed materials. For example, in their study, (11) had the ChatGPT and Google Bard models generate their own questions and analyzed the answers they provided. In contrast, in our study, all multiple-choice questions were prepared by veterinary anatomists, ensuring domain-specific accuracy and minimizing bias in question formulation. This methodological difference underscores that self-generated assessments may overestimate model performance. Had the models been capable of independently generating and reliably answering veterinary anatomy questions, their success rates might have been higher. Therefore, our findings highlight the necessity of expert-driven test design to obtain a realistic evaluation of large language models within specialized fields such as veterinary anatomy. Moreover, because the questions we posed were designed to assess competencies such as the correct use of anatomical terminology and understanding the relationships between structures, it was expected that none of the AI chatbots tested would achieve 100% success. Consequently, although the complete reliability of AI chatbots in the field of veterinary anatomy cannot yet be assured, they may serve as valuable supportive.

Conflict of Interest

The authors declare that there is no conflict of interest related to this study.

Funding

During this study, no financial and/or moral support that could negatively influence the decision to be made regarding the study was received from any pharmaceutical company directly related to the subject of the research, any company supplying and/or manufacturing medical devices, equipment and materials, or any commercial company during the evaluation process of the study.

Authors' Contributions

Motivation / Concept: Ezgi Deniz MAVİLİ, Çağdaş OTO
Design: Aytaç AKÇAY, Çağdaş OTO
Control/Supervision: Çağdaş OTO
Data Collection and / or Processing: Ezgi Deniz MAVİLİ
Analysis and / or Interpretation: Ezgi Deniz MAVİLİ, Aytaç AKÇAY
Literature Review: Ezgi Deniz MAVİLİ, Barış BATUR
Writing the Article: Ezgi Deniz MAVİLİ, Barış BATUR, Aytaç AKÇAY
Critical Review: Çağdaş OTO

Ethical Approval

The data, information and documents presented in this article have been obtained within the framework of academic and ethical standards. Ethical statements have been obtained from the authors, affirming that all information, documents, evaluations, and conclusions are presented in accordance with scientific ethical and moral principles.

REFERENCES

1. Ganapathy A, Kaushal P. Cognitive domain assessment of artificial intelligence chatbots: a comparative study between ChatGPT and Gemini's understanding of anatomy education. Med Sci Educ. 2025;35:1295-1304.
2. Al-Khater KMK. Comparative assessment of three AI platforms in answering USMLE Step 1 anatomy questions or identifying anatomical structures on radiographs. Clin Anat. 2025;38(2):186-199.
3. Choudhary OP, Saini J, Challana A. ChatGPT for veterinary anatomy education: an overview of the prospects and drawbacks. Int J Morphol. 2023;41(4):1198-1202.
4. Arun G, Perumal V, Urias FPJB, Ler YE, Tan BWT, Vallabhajosyula R, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. Anat Sci Educ. 2024;17(7):1396-1405.
5. Singal A, Goyal S. Reliability and efficiency of ChatGPT 3.5 and 4.0 as a tool for scalenovertebral triangle anatomy education. Surg Radiol Anat. 2024;47(1):24.
6. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774; 2023.
7. Bélisle-Pipon JC. Why we need to be careful with large language models in medicine. Front Med. 2024;11:1495582.
8. Bessa RF, de Oliveira AC, Sousa DL, Alves R, Barbosa A, Carneiro A, et al. Performance comparison of large language models on Brazil's medical revalidation exam for foreign-trained graduates. Appl Sci. 2025;15(13):7134.
9. Meo SA, Abukhalaf FA, ElToukhy RA, Sattar K. Exploring the role of DeepSeek-R1, ChatGPT-4, and Google Gemini in medical education: how valid and reliable are they? Pak J Med Sci. 2025;41(7):1887-1892.
10. Campos VMS, Prudente TP, Leão LL, da Costa MS, Oliva HNP, Monteiro-Junior RS. Analyses of different prescriptions for health using artificial intelligence: a critical approach based on international guidelines of health institutions. Health Inf Sci Syst. 2025;13(1):52.
11. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. Cureus. 2023;15(9):e45301.