Two artificial intelligence models underperform on examinations in a veterinary curriculum

Michelle C. Coleman, DVM, PhD, DACVIM*, and James N. Moore, DVM, PhD

Department of Large Animal Medicine and Surgery, College of Veterinary Medicine, University of Georgia, Athens, GA

*Corresponding author: Dr. Coleman (mccole@uga.edu)

OBJECTIVE
Advancements in artificial intelligence (AI) and large language models have rapidly generated new possibilities for education and knowledge dissemination in various domains. Currently, our understanding of the knowledge of these models, such as ChatGPT, in the medical and veterinary sciences is in its nascent stage. Educators are faced with an urgent need to better understand these models in order to unleash student potential, promote responsible use, and align AI models with educational goals and learning objectives. The objectives of this study were to evaluate the knowledge level and consistency of responses of 2 platforms of ChatGPT, namely GPT-3.5 and GPT-4.0.

SAMPLE
A total of 495 multiple-choice and true/false questions from 15 courses used in the assessment of third-year veterinary students at a single veterinary institution were included in this study.

METHODS
The questions were manually entered 3 times into each platform, and answers were recorded. These answers were then compared against those provided by the faculty members coordinating the courses.

RESULTS
GPT-3.5 achieved an overall performance score of 55%, whereas GPT-4.0 had a significantly ($P < .05$) greater performance score of 77%. Importantly, the performance scores of both platforms were significantly ($P < .05$) below that of the veterinary students (86%).

CLINICAL RELEVANCE
Findings of this study suggested that veterinary educators and veterinary students retrieving information from these AI-based platforms should do so with caution.

Keywords: artificial intelligence, education, large language models, ChatGPT, curriculum

Advancements in artificial intelligence (AI) and large language models have rapidly generated new possibilities for education and knowledge dissemination in various domains. In the field of healthcare, where effective clinical decision-making relies on a strong foundation of knowledge and critical thinking skills, incorporating AI-based tools into the learning process has garnered great attention from the public, academics, and science communities.¹

Large language models, such as OpenAI’s ChatGPT, are a form of AI that generate human-like responses and provide interactive conversational experiences. ChatGPT was launched in November of 2022, and within 2 months, it became the fastest-growing consumer application of all time, with over 100 million users.² ChatGPT does not have access to the internet, but rather has been trained on databases from which it extracts and synthesizes relevant information to address a task posed by the user. Recent advances by OpenAI have led to the development of their most advanced language model, ChatGPT 4.0. Currently, ChatGPT 4.0 is trained on a more diverse and extensive database and is only available as a paid subscription, whereas ChatGPT 3.5 is freely available for use by the general public. Both platforms have reached the hands of students in all educational sectors, sparking the urgent need to reevaluate our approach to education.

With over 500 PubMed publications in the 9 months since its launch in 2022, use of ChatGPT in healthcare fields has grown rapidly. Specifically, AI language models have gained popularity in medical education for a variety of applications,³,⁴ including...
information retrieval,5,6 use as a study aid,7,8 clinical decision-making,9,10 medical case discussions,11 and professional development.12 Despite numerous publications in human medicine, there is a paucity of information regarding the use of AI in veterinary medicine and veterinary education, and differences in medical and veterinary-specific training, specifically of large language models, are likely to exist.

Importantly, the use of AI presents several significant limitations, including risk of bias and discrimination, lack of transparency and reliability, the potential for breach in cybersecurity, ethical sequences,15,16 and societal implications.15,16 Furthermore, healthcare provision, whether human or veterinary, requires precision and accuracy, as even small errors may have clinical consequences. Because AI platforms are language models and not knowledge bases, inaccuracy is of great concern.15,16 Today’s educators are faced with the challenges associated with preparing students for the AI age and shaping their learning experiences. Currently, our understanding of the knowledge of ChatGPT in the medical and veterinary sciences is in its nascent stage; however, understanding AI language models is essential so that educators can unleash student potential, promote responsible use, and align AI models with educational goals and learning objectives. Given the limitations of AI and students’ growing use of these platforms, it is critically important that educators and students understand the utility and limitations of this technology. Thus, the objectives of this study were to evaluate the knowledge level and consistency of the performance of ChatGPT-3.5 and -4.0 on examinations used in the veterinary curriculum.

Methods

All examination questions, which were delivered via a digital assessment platform (ExamSoft; ExamSoft Worldwide LLC) to third-year veterinary students at the University of Georgia College of Veterinary Medicine in both core and elective courses during the 2022–2023 academic year, were collected into a database. Each examination was carefully reviewed by both investigators independently to identify questions that would qualify for inclusion in the study. Multiple-choice and true/false questions having a point biserial correlation coefficient (generated by ExamSoft based on student performance) of ≥ 0.15 for item analysis were included. Because ChatGPT only accepts text interactions, questions that included photographs, graphics, charts, or clinical scenarios as image files in the question database that could not be directly entered into the ChatGPT platform were excluded.

Manual input of all questions included the question, the correct answer, and 2 to 4 incorrect options (distractors) for multiple-choice questions. Input was repeated 3 times each into both ChatGPT 3.5 (by MC) and ChatGPT 4.0 (by JM) in July and August of 2023. The same prompt was provided by both evaluators and stated, “Please list in order answers, without explanations, to the following questions. Continue to do this throughout this ChatGPT session.” To prevent error and bias, questions were inputted in sets of up to 5 questions with related questions entered together (Table 1). Repeated entry of each question was spaced by at least 24 hours (ie, each trial was performed on a separate day). All answers were recorded and coded as “correct” or “incorrect” based on the answer provided.

Statistical analysis

The correct answer for each question was provided by the author of each question in ExamSoft. These answers were then compared to the answer provided by ChatGPT. The percentage of correct answers was calculated for each trial. A t test was used to compare the performance of students and each ChatGPT platform and between the 2 ChatGPT platforms. Commercial software (SAS, version 9.4; SAS Institute Inc) was used with significance set at P < .05.

Results

A total of 986 questions were collected from 15 difference courses, including both core (n = 9) and elective (6) courses. Based on inclusion criteria, 495 questions were included in the study. The core courses, taken by all students in the veterinary education program, included Large Animal Theriogenology, Large Animal Digestive Diseases, Musculoskeletal Disease and Pathology, Small Animal Digestive Diseases, Basic Surgical Techniques, Neurology, Cardiology, Respiratory Diseases, and Systemic Pathology II. Elective courses taken by a select number of students included Small Animal Clinical Neurology, Small Animal Oncology, Small Animal Cardiology, Small Animal Advanced Anesthesia, Advanced Topics in Emergency and Critical Care, and Small Animal Musculoskeletal Diseases.

The mean performance score for students for all examinations was 86% (SD, 0.12), with the mean point biserial of included questions being 0.32 (SD, 0.11). The performance scores for each of the 3 trials in GPT-3.5 and GPT-4 are presented in Table 2: the student performance score was significantly greater than those of both GPT-3.5 and GPT-4 (P < .05), and the performance score for GPT-4 was significantly greater than that of GPT-3.5 (P < .05). The consistency in answering a question correctly or incorrectly in all 3 trials for each platform was evaluated. GPT-3.5 answered 33% of questions correctly across all 3 trials and answered 23% of questions incorrectly across all 3 trials. GPT-4 answered 69% of questions correctly across all 3 trials and answered 15% of questions incorrectly across all 3 trials. When all 6 trials are considered (GPT-3.5 and GPT-4.0 combined), ChatGPT was consistently correct on 27% of questions and consistently incorrect on 7% of questions.
Discussion

This study contributes to the understanding of AI’s role in veterinary education and highlights the broader considerations of veterinary educators and students implementing AI tools in the learning process. As AI continues to evolve, both groups must adapt their approaches to align with educational goals and ensure that AI serves as a responsible and effective educational companion.

The accuracy of ChatGPT-3.5 on third-year veterinary examinations at a single institution in this study was 55%, compared to 77% for GPT-4. For both platforms, ChatGPT performed significantly below the level of third-year veterinary students on assessments of their understanding of information across a variety of subjects.

The performance of ChatGPT on several medical examinations has previously been documented. For example, ChatGPT made passing scores on all 3 examinations (Step 1, Step 2, and Step 3) of the US Medical Licensing Examination. However, its level of performance declined as question difficulty increased. The trend of assessing the performance of these platforms extends to specialty examinations such as the Specialty Certificate Examination in dermatology, the Plastic Surgery Inservice Training Examination, and the Ophthalmic Knowledge Assessment Program examination, among others, where ChatGPT demonstrated a satisfactory level of performance. In veterinary medicine, a brief study determined that students and ChatGPT perform similarly on veterinary parasitology examination questions; however, that study was limited to assessing responses to only 3 questions.

Table 1—Template of prompt and questions posed to ChatGPT and response of ChatGPT. Answers may not be correct. These questions are mock examples and are not actual test questions used in the veterinary curriculum.

Prompt: “Please list in order answers, without explanations, to the following questions. Continue to do this throughout this ChatGPT session.”

Which of the following is a common clinical sign in horses with PPID?
A) Increased heart rate
B) Delayed shedding of haircoat
C) Hyperactivity
D) End-expiratory wheezes

What diagnostic test is commonly used to support a diagnosis of PPID in horses?
A) ACTH stimulation test
B) Glucose tolerance test
C) Thyroid function test
D) TRH stimulation test

What medication is commonly used in the treatment of PPID in horses?
A) Dopamine agonist
B) Antimicrobials
C) Antifungals
D) Proton pump inhibitor

What diagnostic test is typically used to diagnose equine asthma?
A) Serum biochemistry
B) Skin biopsy
C) Bronchoalveolar lavage
D) Peritoneal fluid analysis

Which of the following management practices is crucial in reducing respiratory issues in horses with asthma?
A) Providing constant access to hay
B) Increasing exposure to dusty environments
C) Keeping horse in well-ventilated environment
D) Routine dental evaluation

ChatGPT:
B) Delayed shedding of haircoat
A) ACTH stimulation test
A) Dopamine agonist
C) Bronchoalveolar lavage
C) Keeping horse in well-ventilated environment

PPID = Pituitary pars intermedia dysfunction. TRH = Thyrotropin-releasing hormone.

Table 2—Performance score for students and each of the 3 trials in both ChatGPT-3.5 and ChatGPT-4.0

<table>
<thead>
<tr>
<th></th>
<th>GPT-3.5—Trial 1</th>
<th>GPT-3.5—Trial 2</th>
<th>GPT-3.5—Trial 3</th>
<th>GPT-4—Trial 1</th>
<th>GPT-4—Trial 2</th>
<th>GPT-4—Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of correct answers (n = 495)</td>
<td>268</td>
<td>279</td>
<td>272</td>
<td>386</td>
<td>377</td>
<td>381</td>
</tr>
<tr>
<td>No. of incorrect answers (n = 495)</td>
<td>227</td>
<td>216</td>
<td>223</td>
<td>109</td>
<td>118</td>
<td>114</td>
</tr>
<tr>
<td>Average score (%)</td>
<td>54</td>
<td>56</td>
<td>55</td>
<td>78</td>
<td>76</td>
<td>77</td>
</tr>
</tbody>
</table>

The accuracy of ChatGPT-3.5 on third-year veterinary examinations at a single institution in this study was 55%, compared to 77% for GPT-4. For both platforms, ChatGPT performed significantly below the level of third-year veterinary students on assessments of their understanding of information across a variety of subjects.

The performance of ChatGPT on several medical examinations has previously been documented. For example, ChatGPT made passing scores on all 3 examinations (Step 1, Step 2, and Step 3) of the US Medical Licensing Examination. However, its level of performance declined as question difficulty increased. The trend of assessing the performance of these platforms extends to specialty examinations such as the Specialty Certificate Examination in dermatology, the Plastic Surgery Inservice Training Examination, and the Ophthalmic Knowledge Assessment Program examination, among others, where ChatGPT demonstrated a satisfactory level of performance. In veterinary medicine, a brief study determined that students and ChatGPT perform similarly on veterinary parasitology examination questions; however, that study was limited to assessing responses to only 3 questions.
The training data used by large language models, such as ChatGPT, is composed of a diverse range of internet text, from casual conversations to professional articles. The specific frequency of training data on specific topics is not quantified or publicly detailed by OpenAI. Human health is a major area of interest and research, holding a vast internet presence, including medical research, health advice, news about health issues, and personal health experiences. In contrast, while substantial information is available on animal health, it is likely to not be as insidious as human health content. Responses of the model are influenced by the relative representation of these topics in its training data; differences in the accuracy and quality of responses are likely when comparing human health to animal health.

For the purpose of the current study, we elected to compare the performance of ChatGPT-3.5 and ChatGPT-4, both of which were developed to generate text based on input they have received. Importantly, these platforms differ in that ChatGPT-4 contains more computational power and was trained using a more extensive and recent dataset than ChatGPT-3.5, allowing it to address more robust queries covering current concepts. ChatGPT-4 also includes enhancements that allow it to respond to longer prompts than ChatGPT-3.5 and adapt better to specialized domains, which may be relevant to the current project involving information from veterinary medicine. Finally, additional emphasis was employed to reduce biases during the development of ChatGPT-4. Our findings that ChatGPT-4 outperformed ChatGPT-3.5 were similar to the results of a recent report of its performance on the Specialty Certificate Examination in dermatology.

In the current study, only questions having a conservative point biserial correlation coefficient of ≥ 0.15 were included. Point biserial correlation coefficient in the context of examination performance is a statistical measure used to quantify the relationship between the student’s answer to an individual question (correct or incorrect) and their overall performance on an examination. Questions meeting this criterion were considered to have a meaningful relationship between student performance and their ability to answer the question correctly. The performance of students on questions considered to be “less discriminatory” (i.e., point biserial correlation coefficient < 0.15) was not determined.

In the current study, questions that included images were omitted because neither platform has the ability to recognize and interpret images. Clearly, visual information can provide valuable context and visual cues that contribute to the accuracy of assessments and students’ clinical decision-making capabilities. Specifically, images such as radiographs, histopathological slides, and clinical photographs play a crucial role in the pathophysiology, diagnosis, and management of specific conditions. The interest and demand for AI to augment clinical decision-making in human and, more recently, veterinary radiology is growing. As such, large language models, including ChatGPT, may need to incorporate other transformer models to handle multiple types of data. Examples that might be included are contrastive language-image pretraining models that can classify images and generate text-based descriptions.

Prompt engineering refers to the practice of designing, refining, and implementing prompts or instructions to yield effective and optimal large language model output. Due to the significant impact that prompt engineering has on generative output, a standardized input format was utilized in the current study; a different prompt may have generated different responses. Due to the types of questions (i.e., multiple-choice and true/false questions) included in this study, our prompt was designed so that questions could be answered only via a format of “A, B, C, D” or “True, False.” The prompt used in this study allowed for efficient recording of answers and ensured that the session remained focused on providing direct answers to each question. We acknowledge this approach limited our ability to evaluate explanations to answers that would have been provided with a different prompt.

As has been identified by others, AI occasionally provides spurious results. This poorly understood phenomenon, coined artificial hallucination, occurs when AI provides confident answers that cannot be explained by the training data alone. Irrespective of the overall performance, the consistency of results is also important for educators and students to recognize. A key finding of the current study was that ChatGPT-3.5 was correct on all trials for only 33% of questions and ChatGPT-4.0 was correct on all trials for only 69% of questions. Identifying why ChatGPT makes these mistakes may help guide students in interpretation of spurious results.

There were several important limitations to this study. First, all questions were generated by faculty at the University of Georgia for the purpose of assessment of third-year veterinary students. The answers to each question were determined by the faculty member who provided the question at the time the examination was administered. The accuracy of each question was not assessed in this study. A second limitation is that ChatGPT was created from data produced before September 2021. As a result, the platform was trained only on information available before that date. While it is possible that some of the questions included in the current study covered newer information, the take-home message for veterinary students remains the same: do not accept everything ChatGPT reports as being valid.

Another limitation of the current study was that assessment of ChatGPT’s understanding of the material relied on the formulation of answers to multiple choice and true/false questions in a list format, without explanations. As a result, this limited our ability to better appreciate the knowledge base and clinical reasoning of the platform. ChatGPT is constantly updated by OpenAI, and the models used may not represent the latest performance of other AI platforms. Finally, the scope of the current study was limited to questions from a specific veterinary cohort at a single institution, limiting the generalizability of findings to
other veterinary institutions or other educational levels within the veterinary curriculum.

Nonetheless, the findings of the current study provided, to the authors’ knowledge, the first evaluation of the performance of large language models in veterinary education and should caution veterinary students about concluding that AI-generated information is uniformly accurate. The results of this and other studies demonstrate that medical and veterinary students must exercise careful discernment when utilizing information generated by AI language models such as ChatGPT. While these tools can offer valuable insight or information, they lack nuanced understanding, clinical experience, and ethical considerations essential in clinical decision-making. Reliance on AI-generated content carries risk of misinterpretation of information, incorrect diagnoses, and inappropriate treatment planning. As such, it is critically important for students to assess AI-generated information, cross-reference it with reputable sources, consult experienced professionals, and integrate their own expertise prior to making medical decisions. Consequently, AI should be utilized as a supplementary tool rather than as a replacement for human experiences and judgment in the complex and high-stakes fields of medicine and veterinary science.

To further highlight the use of AI technology, the text of this manuscript, excluding the final paragraph above, was entered into ChatGPT-3.5 with the following prompt: “Write a brief conclusion for the following manuscript.” The result is as follows:

“In conclusion, the findings of this study underscore the evolving landscape of education in the AI age. While ChatGPT demonstrates significant promise in enhancing learning experiences in veterinary education, it also underscores the need for ongoing research, critical evaluation, and responsible implementation of AI tools to maximize their benefits and mitigate potential drawbacks.”

Acknowledgments

None reported.

Disclosures

ChatGPT-3.5 and ChatGPT-4.0 were used in the Methods of this manuscript. Description of the manner in which these platforms were used is provided in detail in the Methods section of this paper. Briefly, examination questions were entered 3 times into each platform, and answers to each question were recorded.

Funding

The authors have nothing to disclose.

References


8. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ. 2023;e46885. doi:10.2196/46885


