TYPE Original Research
PUBLISHED 26 August 2025
DOI 10.3389/fvets.2025.1616566

Performance of large language models on veterinary undergraduate multiple-choice examinations: a comparative evaluation

Santiago Alonso Sousa[1]*, Syed Saad Ul Hassan Bukhari[1], Paulo Vinicius Steagall[1,2], Paweł M. Bęczkowski[1], Antonio Giuliano[1] and Kate J. Flay[1]

[1]Department of Veterinary Clinical Sciences, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong SAR, China, [2]Centre for Animal Health and Welfare, City University of Hong Kong, Kowloon, Hong Kong SAR, China

Check for updates

OPEN ACCESS

EDITED BY
Jose Antonio Tapia,
University of Extremadura, Spain

REVIEWED BY
Alasdair James Charles Cook,
University of Surrey, United Kingdom
José Terrado Vicente,
Universidad CEU Cardenal Herrera, Spain

*CORRESPONDENCE
Santiago Alonso Sousa
✉ salonsos@cityu.edu.hk

RECEIVED 23 April 2025
ACCEPTED 14 August 2025
PUBLISHED 26 August 2025

CITATION
Alonso Sousa S, Bukhari SSUH, Steagall PV, Bęczkowski PM, Giuliano A and Flay KJ (2025) Performance of large language models on veterinary undergraduate multiple-choice examinations: a comparative evaluation. *Front. Vet. Sci.* 12:1616566.
doi: 10.3389/fvets.2025.1616566

The integration of artificial intelligence, particularly large language models (LLMs), into veterinary education and practice presents promising opportunities, yet their performance in veterinary-specific contexts remains understudied. This research comparatively evaluated the performance of nine advanced LLMs (ChatGPT o1Pro, ChatGPT 4o, ChatGPT 4.5, Grok 3, Gemini 2, Copilot, DeepSeek R1, Qwen 2.5 Max, and Kimi 1.5) on 250 multiple-choice questions (MCQs) sourced from a veterinary undergraduate final qualifying examination. Questions spanned various species, clinical topics and reasoning stages, and included both text-based and image-based formats. ChatGPT o1Pro and ChatGPT 4.5 achieved the highest overall performance, with correct response rates of 90.4 and 90.8% respectively, demonstrating strong agreement with the gold standard across most categories, while Kimi 1.5 showed the lowest performance at 64.8%. Performance consistently declined with increased question difficulty and was generally lower for image-based than text-based questions. OpenAI models excelled in visual interpretation compared to previous studies. Disparities in performance were observed across specific clinical reasoning stages and veterinary subdomains, highlighting areas for targeted improvement. This study underscores the promising role of LLMs as supportive tools for quality assurance in veterinary assessment design and indicates key factors influencing their performance, including question difficulty, format, and domain-specific training data.

KEYWORDS

veterinary education, artificial intelligence, assessment, large language models, comparative analysis, quality assurance

Introduction

The role of artificial intelligence (AI) in healthcare has become a prominent focus in recent scholarly discussions (1–4). This attention is driven by rapid advancements in large language models (LLMs), a subset of AI systems capable of generating human-like natural language responses from textual input (3). Among current LLMs, ChatGPT, a chat-generative pre-trained transformer developed by OpenAI, has emerged as especially relevant due to its sophisticated deep-learning architecture, trained on extensive datasets, enabling it to produce coherent, contextually appropriate responses to user prompts (5, 6). ChatGPT demonstrates capabilities beyond knowledge recall, with reports of deductive reasoning and chain of thought (CoT) (5).

This facilitates broader applications of ChatGPT (and other comparable AI tools) within the medical field, such as answering medical questions (7), writing medical reports (8), information retrieval (9), aiding study (6, 10) and professional development (11). In 2022, with no specialized training, ChatGPT performed at or near the United States Medical Licensing Exam (USMLE) passing threshold of 60% accuracy (12). Apart from that, ChatGPT has passed general medical licensing examinations from Australia (13), Peru (14) and Iran (15), and has also exceeded passing scores in sub-specialties such as dermatology (16) and radiology (17).

Going forwards, LLMs are predicted to impact all aspects of society, including education and the training and assessment of healthcare professionals (18–20). Recently, more LLMs such as Grok, DeepSeek or Copilot, have shown strong performance in effectively addressing a wide range of queries (21). However, all these LLMs have shown varying proficiency across different healthcare disciplines (21–26), and additional concerns have been raised regarding their understanding of questions, depth of responses and ability to deal with nuanced and context dependent data (2, 11, 27).

Currently, our understanding of the capabilities of LLMs in veterinary science is limited. Few studies have specifically examined the performance of ChatGPT within this context (28–30), and therefore conclusive evidence regarding the accuracy of LLMs in answering veterinary-specific examination questions is lacking. Accuracy in clinical decision-making, whether in human or veterinary medicine, is paramount, as even minor inaccuracies may lead to serious clinical consequences. Given that LLMs primarily function as language-generation systems rather than structured knowledge bases, concerns regarding the accuracy and reliability of their outputs are particularly critical (31, 32). Furthermore, comparative evaluations of LLMs have predominantly focused on human medicine or related healthcare disciplines, with veterinary medicine largely excluded from these assessments (26).

In veterinary education, veterinary graduates are required to demonstrate proficiency across multiple domains, including knowledge, problem-solving, clinical skills, communication, and professionalism (33–35). Although various assessment methods are simultaneously employed to evaluate competence holistically, multiple-choice questions (MCQs) remain extensively utilized as a summative assessment format (36–38). Notably, MCQs constitute the primary format of the North American Veterinary Licensing Examination (NAVLE; International Council for Veterinary Assessment, 2025) (39). Such examinations typically emphasize clinical decision-making, a multifaceted cognitive process that integrates veterinary knowledge, clinical reasoning, the ability to synthesize information from diverse species, and the capacity to apply evidence-based practices.

In this study, we aimed to investigate the potential of LLMs in answering MCQs in veterinary knowledge according to species, subject, type of MCQ (image vs. text-based), clinical reasoning and difficulty levels. The hypothesis was that LLMs would present different accuracy and that would change according to the type of question (species, subject, type of MCQ, clinical reasoning and difficulty levels).

Materials and methods

Large language models

Nine widely recognized LLMs were evaluated in this study between January and February 2025. The models assessed were: (1) ChatGPT o1Pro (OpenAI), (2) ChatGPT 4o (OpenAI), (3) ChatGPT 4.5 (OpenAI), (4) Grok 3 (xAI), (5) Gemini 2 (Google), (6) Copilot (Microsoft), (7) DeepSeek R1 (DeepSeek), (8) Qwen 2.5 Max (Alibaba Cloud), and (9) Kimi 1.5 (Moonshot AI). These specific versions were chosen because they represented the most advanced and updated iterations available during the study period, optimized for reasoning capabilities. Models from OpenAI and xAI were accessed via paid subscription services, whereas the other models were freely accessible to the public at the time of the evaluation.

Multiple-choice questions design and inclusion

The MCQs utilized in this study were derived from the final qualifying examination for the Bachelor of Veterinary Medicine (BVM) Program at the City University of Hong Kong. A total of 250 MCQs were included, each structured as a clinical vignette with a single best answer selected from five options, comprising one correct response and four plausible distractors. These questions were developed to assess knowledge and cognitive skills expected from Day 1 veterinary graduates, aligning with international competency frameworks such as the Australasian Veterinary Boards Council (AVBC) Attributes (35), Royal College of Veterinary Surgeons (RCVS) Day One Competences (33), and the World Organization for Animal Health (WOAH) Competencies of Graduating Veterinarians (40).

The MCQs underwent external benchmarking with peer institutions from the United Kingdom, Australia, and the United States, mirroring the format of the NAVLE. The inclusion and selection criteria for MCQs included reliability, reproducibility, fairness, objectivity, credibility, simplicity of administration, and potential for facilitating constructive feedback.

Content validity was ensured through a comprehensive competency-based educational blueprint, created by three independent expert groups consisting of five subject matter experts each, aligning closely with NAVLE guidelines. This blueprint accounted for the relative importance and instructional hours dedicated to each subject area and animal species within the veterinary curriculum.

MCQ writers, including veterinary specialists and faculty members, received training in best practices for MCQ development. Initial MCQ drafts underwent critical review and individualized feedback from external experts in veterinary education, ensuring adherence to established quality standards. Questions identified as non-compliant were revised and resubmitted until they met the quality criteria. Reliability was investigated through beta-testing by qualified veterinarians (faculty and teaching staff), accompanied by psychometric analysis, including item facility, discrimination indices, and point-biserial correlation, conducted using Speedwell examination software (Speedwell Software Ltd., United Kingdom).

The standard-setting process was conducted using the modified Angoff method (41), where a panel of veterinarians (i.e., instructors) with diverse subject matter expertise, including specialists and general practitioners, were trained on standard-setting principles and inter-rater agreement before establishing the minimum passing score. This criterion-based pass mark ensured fairness and objectivity, independent of candidate group performance. Post-examination psychometric analyses verified the ongoing validity, reliability, and alignment with the veterinary competency framework.

For analytical purposes, MCQs were further classified according to their difficulty levels (low difficulty [≥70% of instructors expect students to answer correctly]), (medium difficulty [40–69% of instructors expect students to answer correctly]), (high difficulty ≤39% of instructors expect students to answer correctly); species (Canine, Feline, Equine, Bovine, Other Production Animals [including small ruminants and swine], and Exotics [including birds, reptiles, amphibians, and small mammals]); image or text-based questions; clinical reasoning stages (e.g., diagnostic interpretation, diagnostic plan, treatment/prognosis, clinical assessment, and prophylaxis); and specific veterinary subdomains (e.g., cardiology, oncology, dentistry, anesthesia/pain management, diagnostic imaging, behavior, and soft tissue surgery) for subsequent analysis.

LLM performance

To ensure that the evaluation was unaffected by prior interactions, each LLM chatbot was assessed using a newly created account with no previous conversation history. A standardized, structured prompt (*"Which of the following is the most appropriate answer for this question: 1, 2, 3, 4 or 5?"*) was consistently inputted into each AI model, followed by clear separation before presenting the MCQs and their corresponding options numbered from 1 to 5. The structured prompt was specifically designed to maintain consistency across chatbot interactions and minimize bias. The MCQs required selecting only the best answer from the given options, without providing any additional justification. Several questions included visual elements (images), which were directly attached as JPG files within the chatbot interface. MCQs were systematically delivered in batches of five to prevent potential chatbot overload or context loss. Each chatbot's responses were documented via manual transcription for subsequent analysis.

Statistical analysis

Responses provided by each chatbot were compared to the gold standard (correct answer sheet), and performance was expressed as agreement percentages (%). Cohen's kappa coefficient (κ) (42, 43) was calculated to assess overall agreement between each chatbot and the gold standard. Further analyses using Cohen's kappa coefficient were conducted to evaluate agreement within specific subcategories, including question difficulty levels (low, medium, and high), species, type of clinical reasoning required, presence or absence of images, and clinical categories. Cohen's kappa values were interpreted following established guidelines: 0–0.20 indicating no agreement, 0.21–0.39 minimal, 0.40–0.59 weak, 0.60–0.79 moderate, 0.80–0.90 strong, and >0.90 almost perfect agreement (42, 43). All statistical analyses were performed using RStudio version 2022.07.1–554 (44).

Results

In general, ChatGPT o1 Pro and ChatGPT 4.5 models had the highest agreement rate, followed by Copilot, DeepSeek R1 and ChatGPT 4o, all showing a strong agreement level. None of the models achieved almost perfect agreement (Cohen's kappa values of >0.90). Kimi 1.5 model performed the worst with a weak level of agreement, and a rate of correct responses of only 64.8% (Table 1).

Agreement rate between LLMs according to level of difficulty

There was almost perfect agreement between all chatbots and the gold standard at low difficulty level (κ = 1.0). At a medium difficulty level, the agreement from ChatGPT o1Pro and ChatGPT 4.5 remained almost perfect; however, agreement was strong to the other chatbots, except Grok 3, Gemini 2 and Kimi 1.5, which presented moderate agreement. At a high-level difficulty ChatGPT o1PRO, ChatGPT 4.5, ChatGPT 4o and Copilot showed moderate agreement whereas Kimi 1.5 presented minimal agreement (Figure 1).

Agreement rate between LLMs based on image-based and non-image-based questions

The agreement between OpenAI chatbots and the gold standard was strong for both text-based and image-based questions. In contrast, Kimi 1.5 showed weak agreement with the gold standard, achieving κ = 0.46 for image-based questions and κ = 0.57 for text-based questions (Table 2). Although the categorical agreement levels for each LLM did not differ between image-based and text-based questions, overall performance tended to be lower on questions containing images.

Agreement rate of chatbots with the gold standard according to species

ChatGPT o1Pro showed almost perfect agreement with "Bovine" and "Other Production Animals," while ChatGPT 4.5 exhibited almost perfect agreement with "Feline." Both chatbots maintained strong agreement levels with the other species. DeepSeek R1 presented strong agreement with four species. In contrast, Kimi 1.5 had the lowest performance, showing minimal agreement with

TABLE 1 Agreement percentages (performance) of different LLMs compared to the gold standard (correct answers) on 250 veterinary MCQs.

Chatbot	Correct answers (%)	Cohen's kappa coefficient
ChatGPT o1 Pro	90.4	0.88
ChatGPT 4o	85.5	0.81
ChatGPT 4.5	90.8	0.88
Grok 3	79.2	0.73
Gemini 2	77.2	0.71
Copilot	85.6	0.82
DeepSeek R1	85.6	0.82
Qwen2.5 Max	79.6	0.74
Kimi 1.5	64.8	0.56

Results are presented as percentages (%) of correct answers and Cohen's kappa coefficients (κ) for overall agreement; 0–0.20 indicating no agreement, 0.21–0.39 minimal, 0.40–0.59 weak, 0.60–0.79 moderate, 0.80–0.90 strong, and >0.90 almost perfect agreement.

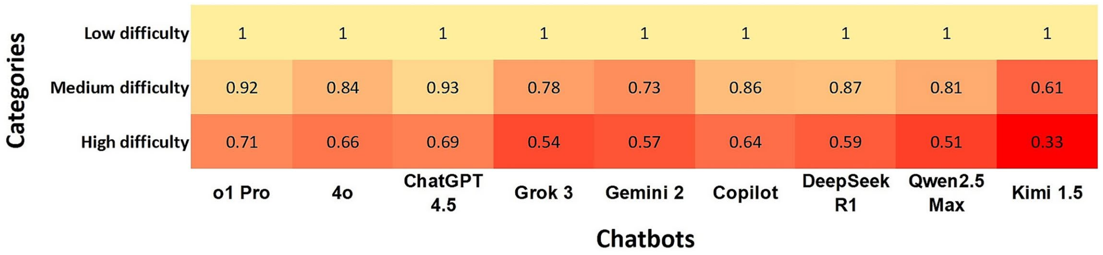


FIGURE 1
Heatmap shows the agreement of chatbots with the gold standard using Cohen's kappa coefficients at questions defined as low, medium, and high difficulty levels. In the heatmap, a darker box color indicates lower agreement, while a lighter color indicates higher agreement.

TABLE 2 Agreement of chatbots with the gold standard for photo and non-photo questions.

Chatbots	Cohen's kappa coefficients for image-based questions	Cohen's kappa coefficients for non-image-based questions
ChatGPT o1 Pro	0.82	0.88
ChatGPT 4o	0.82	0.81
ChatGPT 4.5	0.81	0.89
Grok 3	0.63	0.75
Gemini 2	0.62	0.74
Copilot	0.72	0.79
DeepSeek R1	0.78	0.79
Qwen2.5 Max	0.72	0.74
Kimi 1.5	0.46	0.57

The values in the table represent Cohen's kappa coefficients (κ).

"Exotic" and weak agreement with three species (Figure 2). When comparing species, "Feline" and "Other Production Animals" demonstrated the strongest agreement rates among the chatbots, while "Canine" and "Exotics" categories had the lowest agreement rates (Figure 2).

Agreement rate of chatbots with the gold standard according to clinical reasoning

ChatGPT o1Pro, ChatGPT 4.5, and DeepSeek R1 demonstrated the highest overall agreement rates with the gold standard across all categories, while Kimi 1.5 exhibited the lowest overall agreement. "Diagnostic Interpretation" and "Diagnostic Plan" showed the highest agreement rates among chatbots, whereas "Prophylaxis" had the lowest level of agreement (Figure 3).

Agreement rate of the chatbots with the gold standard according to subdomains

ChatGPT o1Pro and ChatGPT 4.5 had almost perfect agreement in nine and eight subdomains, respectively. Copilot and DeepSeek R1 showed perfect agreement in six subdomains. Conversely, Kimi 1.5 had minimal agreement in six subdomains and non-agreement in one.

When evaluating agreement by subdomains, the highest overall agreement percentages were observed in Endocrinology/Metabolic Diseases, Cardiology, Oncology, and Ophthalmology. In contrast, the lowest agreement levels occurred in Dentistry, Behavior, Anesthesia/Pain Management, Diagnostic Imaging, and Soft Tissue Surgery.

Discussion

In this study, we evaluated the performance of nine LLMs using a dataset of MCQs sourced from the final qualifying examination for the BVM Program at the City University of Hong Kong. To our knowledge, this represents the first comparative evaluation of LLMs targeting veterinary clinical knowledge. Notably, ChatGPT o1Pro and ChatGPT 4.5 outperformed the other chatbots assessed in our study. These results may be explained by OpenAI's recent efforts to enhance deductive reasoning and critical judgment capabilities through advanced prompting techniques such as CoT (45, 46). CoT prompts allow these models to systematically deconstruct complex problems, enhancing their ability to interpret questions accurately and provide precise answers.

DeepSeek R1 achieved agreement rates similar to ChatGPT 4o and Copilot, the latter being based on ChatGPT 4o's architecture, especially considering its relatively recent development. This comparable performance is particularly notable given differences in the underlying architecture and training approaches of these models. OpenAI utilizes reinforcement learning from human feedback (RLHF), a technique where models are fine-tuned using human-generated reward signals to optimize output quality and alignment. This approach is combined with supervised fine-tuning (SFT), where models learn from labeled datasets containing input–output pairs provided by human annotators. These methods are employed in their advanced models during the pre-training phase. In contrast, DeepSeek integrates reinforcement learning and SFT applied to pre-trained data within a mixture-of-experts architecture (47). On the other hand, Kimi 1.5 had weak agreement and the lowest performance when compared with the other LLMs. In previous studies, Kimi 1.5 performed better than ChatGPT 4o and equal to ChatGPT o1 (48) with respect to reasoning; however, our study did not corroborate these findings. This may be due to the build-in model architecture. Kimi's reinforcement learning (RL) approach avoided questions such as MCQs, because these could be answered correctly without good reasoning which could hack the reward model, potentially leading to a possible worse performance with MCQs (48).

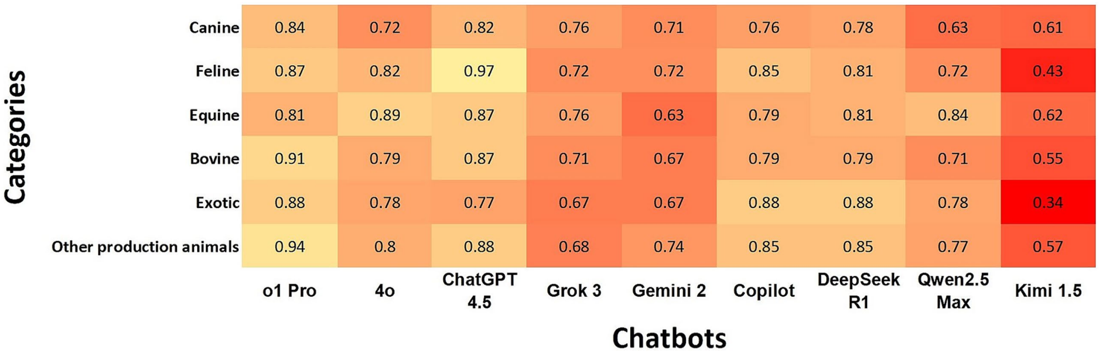


FIGURE 2
Heatmap shows the agreement of chatbots with the gold standard using Cohen's kappa coefficients according to species. In the heatmap, a darker box color indicates lower agreement, while a lighter color indicates higher agreement.

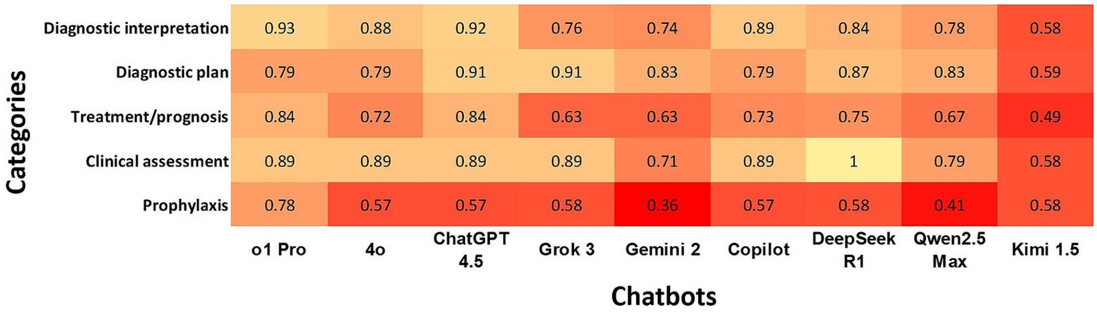


FIGURE 3
Heatmap shows the agreement of chatbots with the gold standard using Cohen's kappa coefficients according to clinical reasoning. In the heatmap, a darker box color indicates lower agreement, while a lighter color indicates higher agreement.

High accuracy of LLMs in veterinary knowledge has become increasingly important as a growing number of pet owners rely on these platforms for guidance regarding their pets' health (49). Although these systems can enhance pet owners' general knowledge and understanding, they also carry significant risks of misinformation, which may lead to detrimental outcomes (50). Our study, which evaluated some of the newest LLMs known for advanced reasoning capabilities, yielded findings consistent with previous research that compared ChatGPT 3.5 and ChatGPT 4 in answering MCQs and true/false questions, reporting accuracy rates of 55 and 77%, respectively (30).

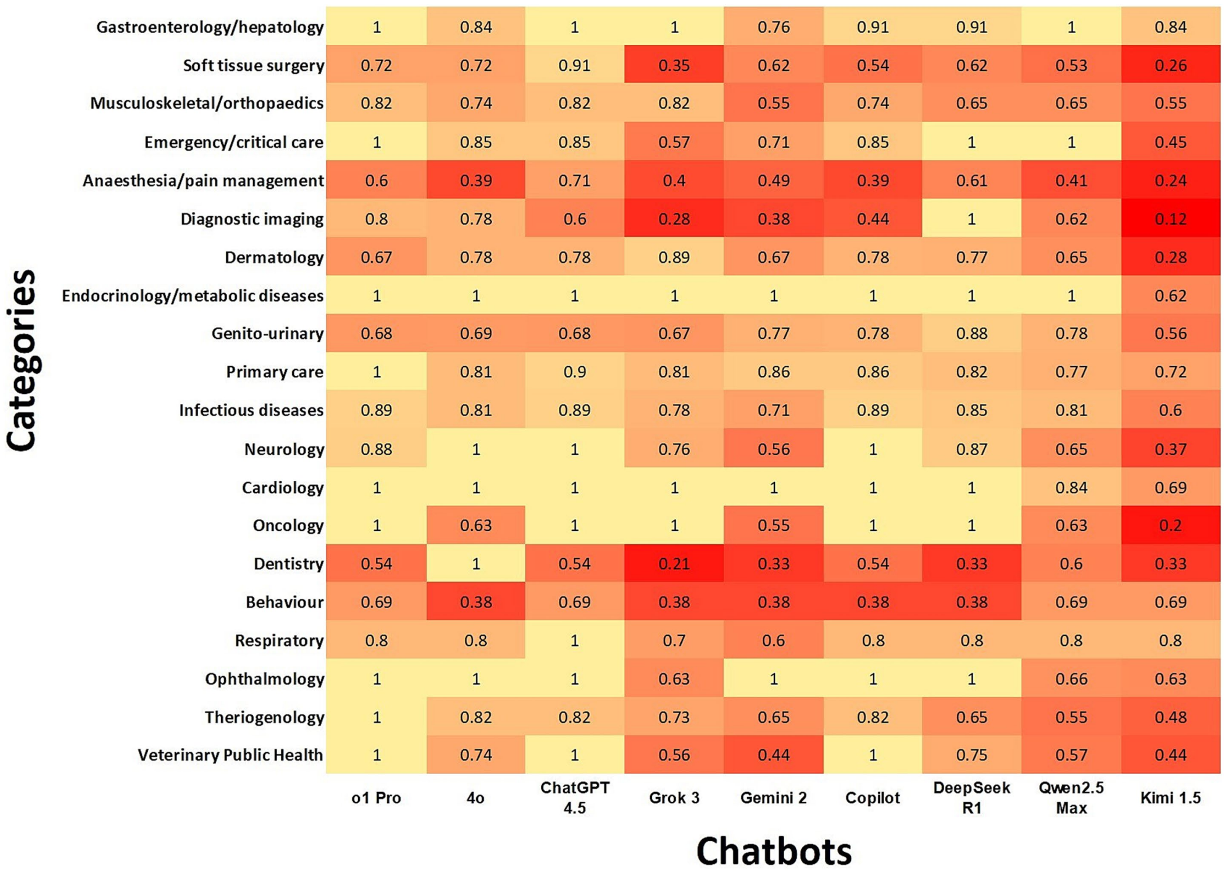
When comparing the agreement levels of different LLMs with the gold standard across low, medium, and high difficulty MCQs, our results indicate that all models achieved almost perfect agreement at low difficulty levels. However, agreement levels decreased as question difficulty increased. ChatGPT o1Pro and ChatGPT 4.5 maintained a moderate level of agreement even at higher difficulty levels. These findings align with previous research showing that ChatGPT's performance on the United States Medical Licensing Examination declined as question difficulty increased (7). Our results also parallel higher LLM performance on the primary-care-level National Certificate Examination for Primary Diabetes Care in China compared to specialist-level Specialty Certificate Examination in Endocrinology and Diabetes administered by the Royal College of Physicians of the United Kingdom (21). This decline in performance at higher difficulty levels is likely due to limitations in the models' logical reasoning capabilities. As questions become more challenging, correct answers depend on domain-specific knowledge and sound logical reasoning capabilities, which can become strained when cognitive demands exceed the model's current training and reasoning capacities (7).

A distinctive aspect of our study is the inclusion of both text-based and image-based questions, reflecting the integral role of visual assessment in veterinary practice. LLMs had similar agreement levels when comparing text and image-based questions. However, there was an overall lower performance with image-based questions using almost all models. OpenAI models were the only LLMs to consistently achieve a strong agreement level across both question types. These findings align with other healthcare studies, which similarly observed decreased performance of LLMs on image-based questions compared to text-based questions (51, 52). However, when comparing our results to previous evaluations of OpenAI models, our study demonstrates a significant improvement. The agreement performance increased from 57% on image-based questions in a study assessing ChatGPT 4 on the Fellowship of the Royal College of Surgeons Trauma and Orthopedics examination (53) to 89% accuracy with OpenAI models in our study. This notable advancement could indicate a substantial progress in these models' capacity to analyze and interpret image-based data, a critical skill, particularly in specialties such as diagnostic imaging. The observed improvement likely results from enhanced training techniques and larger, more diverse visual datasets, enabling AI models to learn effectively from a broader spectrum of visual inputs (54, 55).

While our study highlights the strong performance of multiple LLMs across various veterinary domains, we also observed inconsistencies across clinical reasoning (Figure 3) and subdomains (Figure 4). Specifically, within clinical reasoning, the areas of treatment/prognosis and prophylaxis demonstrated the lowest levels of agreement. Additionally, subdomains including dentistry, behavior, anesthesia/pain management, soft tissue surgery, and diagnostic imaging exhibited comparatively lower agreement levels across all LLMs. Several factors may explain these discrepancies. First, the broad and interdisciplinary nature of these subdomains can introduce ambiguity due to extensive and overlapping content across multiple species. Second, rapid advancements in specialized areas such as anesthesia, pain management, and soft tissue surgery may result in knowledge gaps if the training datasets of these LLMs are not sufficiently up-to-date. Third, decreased performance in diagnostic imaging aligns with our earlier findings that LLMs exhibit inherent limitations in accurately analyzing and interpreting image-based data. Additionally, although LLMs are trained on extensive datasets encompassing various domains, human health represents a significantly larger portion of available data compared to veterinary medicine (30). This disproportionate focus on human medical data could contribute to disparities in accuracy between veterinary and human health responses (30). Finally, it is possible that these questions require additional internal audit. Our findings suggest that despite the demonstrated potential of certain LLMs in veterinary medicine, further targeted training and model development are essential to achieve consistent and reliable performance across all veterinary subjects.

Our evaluation of LLM performance relied on MCQs, a format known to often contain inherent imperfections known as item-writing flaws (IWFs) (56). Although flaws may seem minor, they can influence how learners interpret and respond to questions, potentially leading to misleading results (57, 58). IWFs could partially explain why, out of the 250 MCQs used in our study, there were six questions that all evaluated LLMs answered incorrectly, and 21 questions where more than 50% of the models failed to provide correct answers. Upon subsequent individual assessment by faculty authors, we determined that several of these incorrectly answered questions did indeed contain IWFs, potentially creating ambiguous or misleading scenarios. The occurrence of IWFs in MCQs can often stem from differences in academic training, variation in clinical expertise among educators in question-writing practices (59), or from constraints such as insufficient time to adequately develop high-quality MCQs (60). Thus, enhancing MCQ quality through targeted interventions appears valuable. This challenge also creates an opportunity for integrating LLMs into veterinary education as supportive tools for quality assurance and internal audit, helping educators with the identification and revision of MCQs that warrant additional scrutiny (61).

This study has several limitations. First, we evaluated nine LLMs available at the time of the study. Given the rapid evolution and



	o1 Pro	4o	ChatGPT 4.5	Grok 3	Gemini 2	Copilot	DeepSeek R1	Qwen2.5 Max	Kimi 1.5
Gastroenterology/hepatology	1	0.84	1	1	0.76	0.91	0.91	1	0.84
Soft tissue surgery	0.72	0.72	0.91	0.35	0.62	0.54	0.62	0.53	0.26
Musculoskeletal/orthopaedics	0.82	0.74	0.82	0.82	0.55	0.74	0.65	0.65	0.55
Emergency/critical care	1	0.85	0.85	0.57	0.71	0.85	1	1	0.45
Anaesthesia/pain management	0.6	0.39	0.71	0.4	0.49	0.39	0.61	0.41	0.24
Diagnostic imaging	0.8	0.78	0.6	0.28	0.38	0.44	1	0.62	0.12
Dermatology	0.67	0.78	0.78	0.89	0.67	0.78	0.77	0.65	0.28
Endocrinology/metabolic diseases	1	1	1	1	1	1	1	1	0.62
Genito-urinary	0.68	0.69	0.68	0.67	0.77	0.78	0.88	0.78	0.56
Primary care	1	0.81	0.9	0.81	0.86	0.86	0.82	0.77	0.72
Infectious diseases	0.89	0.81	0.89	0.78	0.71	0.89	0.85	0.81	0.6
Neurology	0.88	1	1	0.76	0.56	1	0.87	0.65	0.37
Cardiology	1	1	1	1	1	1	1	0.84	0.69
Oncology	1	0.63	1	1	0.55	1	1	0.63	0.2
Dentistry	0.54	1	0.54	0.21	0.33	0.54	0.33	0.6	0.33
Behaviour	0.69	0.38	0.69	0.38	0.38	0.38	0.38	0.69	0.69
Respiratory	0.8	0.8	1	0.7	0.6	0.8	0.8	0.8	0.8
Ophthalmology	1	1	1	0.63	1	1	1	0.66	0.63
Theriogenology	1	0.82	0.82	0.73	0.65	0.82	0.65	0.55	0.48
Veterinary Public Health	1	0.74	1	0.56	0.44	1	0.75	0.57	0.44

FIGURE 4
Heatmap illustrates chatbot agreement with the gold standard using Cohen's kappa coefficients according to subdomains. In the heatmap, darker color indicates lower agreement, while lighter color represents higher agreement.

continuous updates of LLM technology, our findings may not fully represent the latest performance of these or other models. Second, our dataset comprised a limited sample size of 250 MCQs, distributed across six veterinary species, potentially insufficient to comprehensively represent the extensive knowledge base of LLMs. Third, we did not have access to the average scores of veterinary students who took the original examination, which prevented direct comparison of student performance with LLMs outcomes. Fourth, although MCQs are widely utilized to assess foundational knowledge, we did not evaluate the reasoning processes underlying the models' decision-making. Consequently, our study does not reflect the models' capabilities in handling complex, open-ended clinical scenarios. Lastly, as indicated in prior research, the reproducibility of answers provided by LLMs has been inconsistent, resulting in variable responses upon repeated questioning (62). Nevertheless, recent studies have noted significant improvements in reproducibility among newer LLMs (63). Furthermore, previous studies have demonstrated that LLM performance significantly decreases when responding to prompts in languages other than English, likely due to the dominance of English-language data in their training datasets (64). Consequently, the findings presented here may not directly extrapolate to multilingual veterinary education settings, highlighting the need for future research to assess LLM accuracy across different languages and to enhance multilingual support for equitable global use.

In conclusion, this comparative evaluation of LLMs highlights their varied strengths and weaknesses across different veterinary domains, with ChatGPT o1Pro and ChatGPT 4.5 demonstrating the strongest overall performance. Most of the evaluated LLMs exhibited improved accuracy compared to previous veterinary-focused studies. Key factors influencing performance included question difficulty, with higher complexity significantly reducing model accuracy, and question format, with image-based questions generally yielding lower performance than text-based ones. These findings highlight the potential role of LLMs as valuable supportive tools in veterinary education, particularly for quality assurance in assessment design and implementation.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: confidentiality. Requests to access these datasets should be directed to salonsos@cityu.edu.hk.

Author contributions

SAS: Data curation, Conceptualization, Project administration, Formal analysis, Writing – review & editing, Writing – original draft. SSUHB: Formal analysis, Visualization, Writing – review & editing. PS: Supervision, Writing – review & editing. PB: Writing – original draft, Methodology, Writing – review & editing. AG: Methodology, Writing – review & editing. KF: Writing – review & editing, Supervision, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health.* (2023) 5:e105–6. doi: 10.1016/S2589-7500(23)00019-5

2. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digit Health.* (2023) 2:e0000205. doi: 10.1371/journal.pdig.0000205

3. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. *Clin Med (Lond).* (2023) 23:278–9. doi: 10.7861/clinmed.2023-0078

4. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns.* (2024) 5:100943. doi: 10.1016/j.patter.2024.100943

5. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiol.* (2023) 1:100017. doi: 10.1016/j.metrad.2023.100017

6. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ.* (2024) 17:926–31. doi: 10.1002/ase.2270

7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* (2023) 9:e45312. doi: 10.2196/45312

8. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol.* (2024) 34:2817–25. doi: 10.1007/s00330-023-10213-1

9. Rao SJ, Isath A, Krishnan P, Tangsrivimol JA, Virk HUH, Wang Z, et al. ChatGPT: a conceptual review of applications and utility in the field of medicine. *J Med Syst.* (2024) 48:59. doi: 10.1007/s10916-024-02075-x

10. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ.* (2019) 5:e13930. doi: 10.2196/13930

11. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* (2023) 6:1169595. doi: 10.3389/frai.2023.1169595

12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. (2023) 2:e0000198. doi: 10.1371/journal.pdig.0000198

13. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and new Bing in an Australian medical licensing examination. *Med J Aust*. (2023) 219:237. doi: 10.5694/mja2.51932

14. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, de la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. *JMIR Med Educ*. (2023) 9:e48039. doi: 10.2196/48039

15. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. *BMJ Health Care Inform*. (2023) 30:e100815. doi: 10.1136/bmjhci-2023-100815

16. Lewandowski M, Łukowicz P, Świetlik D. An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the dermatology specialty certificate examinations. *Clin Exp Dermatol*. (2023) 49:686–691. doi: 10.1093/ced/llad255

17. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. (2023) 307:e230582. doi: 10.1148/radiol.230582

18. Heng JJY, Teo DB, Tan LF. The impact of chat generative pre-trained transformer (ChatGPT) on medical education. *Postgrad Med J*. (2023) 99:1125–7. doi: 10.1093/postmj/qgad058

19. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? *Med Educ Online*. (2023) 28:2181052. doi: 10.1080/10872981.2023.2181052

20. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ*. (2023) 9:e48785. doi: 10.2196/48785

21. Li H, Jiang Z, Guan Z, Liu R, Bao Y, Liu Y, et al. Large language models for diabetes training: a prospective study. *Sci Bull*. (2025) 70:934–42. doi: 10.1016/j.scib.2025.01.034

22. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *J Periodontol*. (2024) 95:682–7. doi: 10.1002/JPER.23-0514

23. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. (2023) 104:269–73. doi: 10.4174/astr.2023.104.5.269

24. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol*. (2023) 13:1256459. doi: 10.3389/fonc.2023.1256459

25. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. (2023) 29:721–32. doi: 10.3350/cmh.2023.0089

26. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J Med Internet Res*. (2024) 26:e56532. doi: 10.2196/56532

27. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. (2024) 151:104620. doi: 10.1016/j.jbi.2024.104620

28. Alonso Sousa S, Flay KJ. A survey of veterinary student perceptions on integrating ChatGPT in veterinary education through AI-driven exercises. *J Vet Med Educ*. (2024):e20240075. doi: 10.3138/jvme-2024-0075

29. Chu CP. ChatGPT in veterinary medicine: a practical guidance of generative artificial intelligence in clinics, education, and research. *Front Vet Sci*. (2024) 11:1395934. doi: 10.3389/fvets.2024.1395934

30. Coleman MC, Moore JN. Two artificial intelligence models underperform on examinations in a veterinary curriculum. *J Am Vet Med Assoc*. (2024) 262:692–7. doi: 10.2460/javma.23.12.0666

31. 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) 9:e46885. doi: 10.2196/46885

32. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the chat-GPT model. *Res Sq*. (2023). doi: 10.21203/rs.3.rs-2566942/v1

33. Royal College of veterinary surgeons (RCVS). Day one competences. London: RCVS (2022).

34. Association of American Veterinary Medical Colleges (AAVMC). Competency-based veterinary education: competency framework. *AAVMC*. (2024)

35. Australasian Veterinary Boards Council (AVBC). Day One Competencies. Melbourne: AVBC (2024).

36. Cook AK, Lidbury JA, Creevy KE, Heseltine JC, Marsilio S, Catchpole B, et al. Multiple-choice questions in small animal medicine: an analysis of cognitive level and structural reliability, and the impact of these characteristics on student performance. *J Vet Med Educ*. (2020) 47:497–505. doi: 10.3138/jvme.0918-116r

37. Baillie S, Rhind S, Warman S. A guide to assessment in veterinary medical education. *3rd* ed. Bristol: University of Bristol (2022).

38. Richter R, Tipold A, Schaper E. Measures for quality assurance of electronic examinations in a veterinary medical curriculum. *J Vet Med Educ*. (2024) 51:762–76. doi: 10.3138/jvme-2023-0061

39. International Council for Veterinary Assessment. North American Veterinary Licensing Examination (NAVLE). Available at: https://www.icva.net/navle/ (Accessed April 1, 2025).

40. World Organisation for Animal Health. (2012). OIE recommendations on the competencies of graduating veterinarians ('Day 1 graduates') to assure national veterinary services of quality. Available at: https://www.woah.org/app/uploads/2021/03/dayone-b-ang-vc.pdf (Accessed April 1, 2025).

41. Hambleton R. Setting performance standards on educational assessments and criteria for evaluating the process In: G Cizek, editor. Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates (2001). 89–115.

42. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer*. (2023) 23:799–5. doi: 10.1186/s12885-023-11325-z

43. McHugh ML. Lessons in biostatistics: interrater reliability—the kappa statistic. *Biochem Med Zagreb*. (2012) 22:276–82. doi: 10.11613/BM.2012.031

44. Posit team. RStudio: Integrated Development Environment for R. Posit Software, PBC (2020). Available at: https://www.rstudio.com/ (Accessed April 1, 2025).

45. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Proces Syst*. (2022) 35:24824–37. doi: 10.48550/arXiv.2201.11903

46. Feng G, Zhang B, Gu Y, Ye H, He D, Wang L. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Adv Neural Inf Proces Syst*. (2023) 36:7075798. doi: 10.48550/arXiv.2305.15408

47. Egger J, Faray de Paiva L, Luijten G, Krittanawong C, Keyl J, Sallam M, et al. Is deep seek-R1 a game changer in healthcare? A seed review. *Preprint*. (2025). doi: 10.13140/RG.2.2.21706.89283

48. Kimi Team. KIMI K1.5: Scaling reinforcement learning with LLMs. *arXiv*. (2025)

49. Jokar M, Abdous A, Rahmanian V. AI chatbots in pet health care: opportunities and challenges for owners. *Vet Med Sci*. (2024) 10:e1464. doi: 10.1002/vms3.1464

50. Souza GV, Hespanha ACV, Paz BF, Sá MAR, Carneiro RK, Guaita SAM, et al. Impact of the internet on veterinary surgery. Vet Anim Sci (2020) 11:100161–1. doi: 10.1016/j.vas.2020.100161

51. Chen CJ, Sobol K, Hickey C, Raphael J. The comparative performance of large language models on the hand surgery self-assessment examination. *Hand*. (2024):1–5. doi: 10.1177/15589447241279460

52. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch Clin Exp Ophthalmol*. (2025) 263:527–36. doi: 10.1007/s00417-024-06625-4

53. Khan AM, Sarraf KM, Simpson AI. Enhancements in artificial intelligence for medical examinations: a leap from ChatGPT 3.5 to ChatGPT 4.0 in the FRCS trauma & orthopaedics examination. *Surgeon*. (2025) 23:13–7. doi: 10.1016/j.surge.2024.11.008

54. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR)*; (2016); Las Vegas, NV, USA. p. 770–778.

55. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst*. (2020) 34:159

56. Przymuszała P, Piotrowska K, Lipski D, Marciniak R, Cerbin-Koczorowska M. Guidelines on writing multiple choice questions: a well-received and effective faculty development intervention. *SAGE Open*. (2020) 10:1–12. doi: 10.1177/2158244020947432

57. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med*. (2002) 77:156–61. doi: 10.1097/00001888-200202000-00016

58. Tariq S, Tariq S, Maqsood S, Jawed S, Baig M. Evaluation of cognitive levels and item writing flaws in medical pharmacology internal assessment examinations. *Pak J Med Sci*. (2017) 33:866–70. doi: 10.12669/pjms.334.12887

59. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ*. (2008) 42:198–206. doi: 10.1111/j.1365-2923.2007.02957.x

60. Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. *J Nurses Prof Dev*. (2013) 29:52–7. doi: 10.1097/NND.0b013e318286c2f1

61. Willert N, Würz PK. Assisting quality assurance of examination tasks: using a GPT model and Bayesian testing for formative assessment. *Comput Educ Artif Intell*. (2025) 8:100343. doi: 10.1016/j.caeai.2024.100343

62. Morishita M, Fukuda H, Muraoka K, Nakamura T, Yoshioka I, Ono K, et al. Comparison of the performance on the Japanese National Dental Examination using GPT-3.5 and GPT-4. *Jpn J Dent Educ Assoc*. (2024) 40:3–10. doi: 10.24744/jdea.40.13

63. Wulcan JM, Jacques KL, Lee MA, Kovacs SL, Dausend N, Prince LE, et al. Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. *Front Vet Sci*. (2025) 11:1490030. doi: 10.3389/fvets.2024.1490030

64. Vadlapati P. Multilingual prompting in LLMs: investigating the accuracy and performance. *Int J Scientific Res Eng Manag (IJSREM)*. (2023) 7. doi: 10.55041/IJSREM17694