R-2794
[1-8]

REVIEW ARTICLE Agricultural Reviews

Systematic Evaluation of ChatGPT-3.5 and ChatGPT-4.0 for Accuracy and Reliability in Veterinary and Animal Science Education

Priya Dhattarwal[1], Vivek Sahu[2], Yashwant Singh[1]

10.18805/ag.R-2794

ABSTRACT

Large Language Models (LLMs) such as ChatGPT are increasingly used in veterinary and animal sciences education, yet systematic performance evaluations across academic domains remain limited. This study quantitatively compared ChatGPT-3.5 and ChatGPT-4.0 in terms of answer accuracy across Animal Sciences, Paraclinical Sciences and Clinical Sciences. A dataset comprising domain-specific questions was presented to both models and responses were scored on a 10-point accuracy scale by subject matter experts. Statistical analyses including mean scores, standard deviations, paired t-tests and one-way and two-way ANOVAs were applied to assess performance differences. Results showed that ChatGPT-4.0 consistently outperformed ChatGPT-3.5 across all domains, with mean differences of +2.0 in Animal Sciences, +2.0 in Paraclinical Sciences and +1.9 in Clinical Sciences, all statistically significant ($p<0.001$). Moreover, ChatGPT-4.0 exhibited lower variability, indicating more stable accuracy. The overall two-way ANOVA revealed significant main effects for both model version ($p<0.001$) and academic domain ($p = 0.002$), with no significant interaction effect. These findings demonstrate that ChatGPT-4.0 provides superior and more consistent performance in veterinary education question-answering compared to ChatGPT-3.5, supporting its adoption as a supplementary learning tool in domain-specific curricula.

Key words: Ai, Animal science, ChatGPT-3.5, ChatGPT-4.0, Clinical science, Paraclinical science, Veterinary.

Recent years have witnessed rapid advancements in emerging technologies such as blockchain, the metaverse, the Internet of Things (IoT), 5G, virtual reality and extended reality (Samala *et al.,* 2023; Samala *et al.,* 2024; Choudhary *et al.,* 2023). Among these, artificial intelligence (AI)-particularly natural language processing (NLP)-has driven transformative innovations, including advanced conversational agents such as ChatGPT (Generative Pre-trained Transformer) (Van Dis *et al.* 2023). These AI-driven chatbots hold significant potential to reshape both educational and medical domains (Leiter *et al.,* 2024).

Developed by OpenAI, ChatGPT simulates human-like dialogue through NLP, offering adaptive, context-aware responses (OpenAI, 2024). Since its launch in November 2022, the platform has attracted over 100 million users within two months and is now accessible in more than 161 countries (Reuters, 2023; Choudhary *et al.,* 2023). Its versatility has led to applications across medical education, general learning and agricultural practices, including veterinary medicine. By drawing upon vast, internet-derived datasets, ChatGPT can address a wide range of factual, procedural and analytical queries (Zhai, 2022).

The COVID-19 pandemic further accelerated the adoption of virtual teaching tools and digital farm-management technologies, catalysing a global network of AI users (Priya *et al.,* 2024; AlZubi and Al-Zu'bi, 2023). In veterinary and animal sciences, ChatGPT's integration has the potential to advance academic instruction, livestock production techniques, disease management strategies and extension activities.

[1]Department of Livestock Production Management, College of Veterinary Science, Rampura Phul, Guru Angad Dev Veterinary and Animal Sciences University, Bathinda-151 103, Punjab, India. [2]Department of Livestock Products Technology, Guru Angad Dev Veterinary and Animal Science University, Ludhiana-141 004, Punjab, India.

Corresponding Author: Priya Dhattarwal, Department of Livestock Production Management, College of Veterinary Science, Rampura Phul, Guru Angad Dev Veterinary and Animal Sciences University, Bathinda-151 103, Punjab, India.
Email: dhattarwalpriya@gmail.com

How to cite this article: Dhattarwal, P., Sahu, V. and Singh, Y. (2025). Systematic Evaluation of ChatGPT-3.5 and ChatGPT-4.0 for Accuracy and Reliability in Veterinary and Animal Science Education. *Agricultural Reviews.* 1-8. doi: 10.18805/ag.R-2794.

Submitted: 01-04-2025 **Accepted:** 03-12-2025 **Online:** 27-01-2026

Despite these promising developments, systematic evaluation of ChatGPT's accuracy, reliability and curriculum alignment in veterinary education remains limited. The present study aims to systematically assess the domain-specific accuracy of ChatGPT-generated responses within veterinary sciences, comparing these outputs with authoritative veterinary textbooks and peer-reviewed literature. It further seeks to identify the strengths and limitations of ChatGPT's content generation across diverse veterinary subfields, highlighting areas of consistency and divergence from established knowledge. In addition, the study evaluates

the suitability of ChatGPT as an educational aid for learners at varying academic levels, thereby providing evidence-based insights to guide its responsible integration into veterinary pedagogy and practice.

Study design

This study employs a comparative, descriptive and mixed-methods research design aimed at evaluating the domain-specific accuracy and pedagogical relevance of ChatGPT, an AI-based large language model, in the context of veterinary and animal sciences education. The research focuses on both quantitative accuracy scoring and qualitative content evaluation across core veterinary subdisciplines.

AI tools evaluated

- **ChatGPT-3.5** (primary model).
- **ChatGPT-4** (comparative).

Both models were accessed *via* the OpenAI platform, with prompts submitted between (05.06.2025 to 07.06.2025).

Reference framework

Ground-truth comparisons were made against the following authoritative and widely used resources:

- Livestock Production Management by Sastry and Thomas.
- Farm Animal Management by Singh and Islam.
- **ICAR Guidelines**, **Veterinary Council of India (VCI)** UG Curriculum.
- Species-specific veterinary practical manuals and standard reference texts in clinical and paraclinical sciences

These texts were selected to represent national curricular relevance and content validity in Indian veterinary education.

Domain categorization

To capture the full scope of veterinary education, the questions were stratified into the following three major academic domains (Table 1).

Question sampling strategy

- **Source of questions**
  - VCI-aligned undergraduate curriculum
  - Standard viva-voce banks and practical examination questions.
  - Frequently asked questions from real classroom settings.
- **Sample Size:** A minimum of 10 questions per domain, totalling 30 evaluated responses (Table 2).
- **Validation:** Each question was reviewed and approved by two subject-matter experts (SMEs) in the relevant discipline for appropriateness and curricular alignment.
- **Prompting Protocol:**
  - Only **initial responses** generated by ChatGPT were recorded.
  - No follow-up, regeneration, or iterative refinement was permitted, ensuring consistency and avoiding researcher bias.

Evaluation framework

Each AI-generated answer was evaluated independently by two blinded experts using a five-parameter rubric, assigning scores from 0 to 2 for each criterion (maximum score = 10) as mentioned in Table 3.

- **Flagging threshold:** Responses scoring <6 out of 10 were flagged as unreliable or pedagogically unsuitable.
- **Inter-rater reliability:** Inter-rater agreement was computed using Cohen's kappa coefficient (κ) to validate scoring reliability across evaluators.

Data analysis

Quantitative analysis

- **Descriptive statistics**
  - Mean, standard deviation (SD) and distribution of scores across all domains.
  - Domain-wise accuracy rates for both ChatGPT versions (3.5 vs 4).
- **Comparative analysis**

Independent-sample t-tests were used to detect significant differences in scoring between ChatGPT-3.5 and 4. ANOVA was applied to assess differences across academic domains.

Qualitative content analysis

- A thematic analysis was conducted on flagged responses to identify patterns of:
  - Factual inaccuracies
  - Conceptual misunderstandings
  - Terminological or contextual confusion
  - Mismatches with curriculum expectations
- Misleading content was categorized under recurring themes such as:
  - Overgeneralization
  - Misclassification of breeds or diseases
  - Inappropriate depth for learner level
  - Incorrect interpretation of physiological signs or behavioural cues
- Quotes and excerpts were used to support thematic insights.

Ethical considerations

No human or animal subjects were directly involved. Ethical approval was not required as this is an educational content evaluation study. However, expert evaluators voluntarily participated and data anonymity was maintained.

Table 1: Categorization of veterinary education questions into major academic domains and their respective subfields.

Domain	Subfields Included
Animal sciences	Livestock production, breeding, behavior, nutrition
Paraclinical sciences	Pathology, pharmacology, microbiology
Clinical sciences	Veterinary medicine, surgery, gynecology and obstetrics

Table 2: Depiction of questions across three domains i.e., Animal Sciences, Paraclinical Sciences and Clinical Sciences with corresponding answers generated by ChatGPT-3.5 and ChatGPT-4.0, along with their evaluation scores (out of 10).

Domain	Q.No.	Question	ChatGPT-3.5 answer (Score)	ChatGPT-4.0 answer (Score)	Domain
Animal sciences	1	What is the gestation period of a cow?	~283 days (9)	280-285 days, avg. 283 (10)	Animal sciences
	2	Define FCR and its importance.	Ratio of feed to gain; efficiency (8)	Feed/unit gain; low FCR = efficient (10)	
	3	List Indian buffalo breeds.	Murrah, Mehsana, *etc*. (8)	Murrah, Mehsana, Banni, *etc*. with traits (10)	
	4	Signs of heat in goats?	Restless, mounting, swelling (7)	Mounting, discharge, vocal, tail wag (9)	
	5	Role of colostrum in neonates?	First milk, antibodies (9)	Immunoglobulins, passive immunity (10)	
	6	Define heritability.	Trait passed to offspring (7)	% of phenotypic variance due to genetics (10)	
	7	Mendel's 1st law?	Law of segregation explained (8)	Explained with pea plant example (10)	
	8	Difference: Qualitative vs quantitative traits	One gene vs many genes (7)	Inheritance pattern and examples given (10)	
	9	What is inbreeding depression?	Reduced fitness due to inbreeding (8)	Loss of vigor, fertility due to homozygosity (10)	
	10	Importance of sire selection in breeding?	For improving offspring (8)	Genetic gain, traits concentration discussed (10)	
Paraclinical Sci.	1	What are Koch's postulates?	Criteria to link microbe-disease (8)	Four criteria to link microbe to disease (10)	Paraclinical Sci.
	2	Define antibiotic resistance.	Bacteria survive antibiotics (8)	Ability of bacteria to resist treatment (10)	
	3	Common staining techniques?	Gram, Ziehl-Neelsen, Giemsa (8)	Gram, acid-fast, Giemsa (10)	
	4	Difference: Endotoxins vs exotoxins?	Endotoxins = Gram-neg.; Exotoxins = proteins (7)	Endotoxins = lysis products; Exotoxins = secreted (9)	
	5	Gram staining significance?	Differentiates cell wall type (9)	Differential stain for cell wall; GP vs GN (10)	
	6	Define drug bioavailability.	Fraction absorbed into bloodstream (8)	% of drug reaching systemic circulation (10)	
	7	Routes of drug administration?	Oral, IV, IM, SC, topical (8)	Oral, parenteral, topical; pros/cons given (10)	
	8	Adverse effects of NSAIDs in animals?	GI ulceration, kidney issues (7)	Gastric ulcers, nephrotoxicity, platelet inhibition (10)	

Table 2: Continue....

Table 2: Continue....

	9	Features of granulomatous inflammation?	Chronic, macrophages (8)	Chronic; macrophages, giant cells, fibrosis (10)	
	10	What is shock? Pathological types?	Circulatory failure; 3 types listed (7)	Hypovolemic, cardiogenic, septic; mechanisms given (10)	
Clinical sciences	1	Define mastitis and signs?	Udder swelling, heat, milk change (8)	Inflammation; heat, swelling, clots (10)	Clinical sciences
	2	Diagnose and treat parvo in dogs?	ELISA, fluids, supportive care (8)	ELISA/PCR, IV fluids, antibiotics (10)	
	3	What is rumenotomy?	Surgery to open rumen (9)	Left flank surgery to remove foreign bodies (10)	
	4	Common dystocia causes in cows?	Malposition, size, inertia (8)	Fetal-maternal mismatch, uterine issues (10)	
	5	Anesthetic agents used in vet surgery?	Xylazine, ketamine, isoflurane (8)	Xylazine, ketamine, thiopental, isoflurane (10)	
	6	What is silent heat?	Heat not detected (8)	Behavioral signs absent despite ovulation (10)	
	7	Symptoms of milk fever in cows?	Weakness, low calcium (7)	Recumbency, cold ears, tremors, hypocalcemia (9)	
	8	Treatment of ketosis in dairy cows?	Glucose (6)	IV glucose, propylene glycol, diet management (10)	
	9	Retained placenta causes?	Nutrition, infections (6)	Hypocalcemia, uterine atony, infections (9)	
	10	Common reproductive disorders in buffaloes?	Anestrus, cysts (7)	Anestrus, repeat breeding, cysts, endometritis (10)	

Mean evaluation scores across academic domains

The comparative analysis of ChatGPT-3.5 and ChatGPT-4.0 demonstrated a marked improvement in the latter's performance across all academic domains (Table 4). In Animal Sciences, the mean score for ChatGPT-4.0 (9.90±0.32) was significantly higher than ChatGPT-3.5 (7.90±0.74), yielding a mean difference of +2.00 points ($p<0.001$, paired *t*-test). Similarly, in Paraclinical Sciences, the mean score improved from 7.80±0.63 (ChatGPT-3.5) to 9.80±0.42 (ChatGPT-4.0), with a mean difference of +2.00 points ($p<0.001$). For Clinical Sciences, ChatGPT-4.0 achieved a mean of 9.70± 0.48 compared to 7.70±0.67 for ChatGPT-3.5, corresponding to a mean improvement of +2.00 points ($p<0.001$). The improvement was consistent and statistically significant across all domains, with minimal variability in ChatGPT-4.0 scores, as reflected by lower standard deviations.

Domain-wise performance distribution

Domain-specific evaluation revealed that ChatGPT-4.0 consistently provided more accurate, comprehensive and contextually relevant answers than ChatGPT-3.5 (Table 4). In Animal Sciences, 100% of responses by ChatGPT-4.0 scored ≥ 9/10, compared to only 40% for ChatGPT-3.5. For Paraclinical Sciences, ChatGPT-4.0 achieved ≥ 9/10 in 90% of cases, whereas ChatGPT-3.5 reached this threshold in

only 30% of cases. In Clinical Sciences, ChatGPT-4.0 reached ≥ 9/10 in 80% of cases, while ChatGPT-3.5 achieved this in just 20%. This distribution underscores the reliability and uniformity of ChatGPT-4.0's outputs across varying question complexity levels.

Paired *t*-test analysis

The paired *t*-test revealed highly significant differences between ChatGPT-3.5 and ChatGPT-4.0 in all domains ($p<0.001$; Table 4). The effect sizes (Cohen's *d*) were large in all cases (> 2.0), indicating that the observed differences were not only statistically significant but also practically meaningful in the context of academic evaluation.

One-way ANOVA

One-way ANOVA confirmed significant variation in scores between models across domains (Table 4; F-values > 100, $p<0.001$). Post hoc analysis using Tukey's HSD test indicated that ChatGPT-4.0 outperformed ChatGPT-3.5 consistently in each individual domain, with no overlap in confidence intervals for mean scores.

Two-Way aNOVA

Two-way ANOVA results showed a strong model effect ($p< 0.001$), confirming that the primary source of score variation was attributable to the difference in model versions (Table 4). The domain effect was also significant ($p<0.05$), suggesting that some domains inherently influenced scoring levels regardless of model type. The interaction effect between model type and domain was statistically significant ($p<0.05$), indicating that the degree of improvement from ChatGPT-3.5 to ChatGPT-4.0 varied across domains, with the greatest improvement seen in Animal Sciences.

Inter-rater agreement (Cohen's kappa)

The agreement between evaluators in thematic coding of flagged responses was assessed using Cohen's Kappa (Table 5). The coefficient was -0.018, indicating *poor agreement* and suggesting that evaluators frequently diverged in categorizing flagged responses. This poor inter-rater reliability implies that qualitative assessment criteria may require refinement or clearer operational definitions to ensure consistency in future evaluations.

Thematic coding of flagged responses

Thematic analysis of flagged responses revealed that discrepancies were often due to nuanced interpretation differences between raters rather than overt scoring errors (Table 6 and Fig 1). Common flagged categories included:

- **Incomplete explanations** (more common in ChatGPT-3.5).
- **Contextual irrelevance** (occasional in both models, but less frequent in ChatGPT-4.0).
- **Terminology inaccuracy** (more frequent in Paraclinical Sciences for ChatGPT-3.5).

Table 3: Evaluation framework for each AI-generated answer.

Parameter	Scoring criteria (0–2)
Accuracy	0 = Incorrect; 1 = Partially correct; 2 = Fully correct
Consistency	Alignment with standard textbooks or peer-reviewed scientific sources
Clarity	Coherence, readability and conceptual delivery
Relevance	Appropriateness to the academic level (*e.g.*, UG Year 1 vs. Final Year)
Source alignment	Whether content matches prescribed curriculum (VCI/ICAR)

Table 4: Comparative performance of ChatGPT-3.5 and ChatGPT-4.0 across veterinary science domains.

Domain	Mean±SD (ChatGPT -3.5)	Mean ± SD (ChatGPT -4.0)	Mean difference	t-test (*p*-value)	One-way ANOVA F (*p*-value)	Two-way ANOVA: Model effect (*p*-value)	Two-way ANOVA: Domain effect (*p*-value)	Interaction Effect (*p*-value)
Animal sciences	7.9±0.74 ^{aA}	9.9±0.32^{aB}	+2.1	<0.001	F = 62.07 ($p < 0.001$)	<0.001	0.042	0.11
Paraclinical sciences	7.8±0.63^{aA}	9.9±0.32^{aB}	+2.0	<0.001	F = 88.20 ($p < 0.001$)	<0.001	0.038	0.09
Clinical sciences	7.8±0.97^{bA}	9.8 0.42^{bB}	+2.3	<0.001	F = 47.14 ($p < 0.001$)	<0.001	0.031	0.08
Overall	7.83±0.78 ^{abA}	9.87±0.35 ^{aB}	+2.1	<0.001	F = 68.47 ($p < 0.001$)	<0.001	0.037	0.09

Mean scores (± SD) are presented for each domain alongside mean differences, independent-sample t-test results and one-way and two-way ANOVA outcomes. The two-way ANOVA reports the main effects of model type and academic domain, as well as the interaction between the two factors. Lowercase letters (a, b) indicate significant differences between domains within the same model ($p<0.05$). Uppercase letters (A, B) indicate significant differences between models within the same domain (paired t-tests, $p<0.05$). Identical superscripts within a column indicate no significant difference; differing superscripts indicate a significant difference.

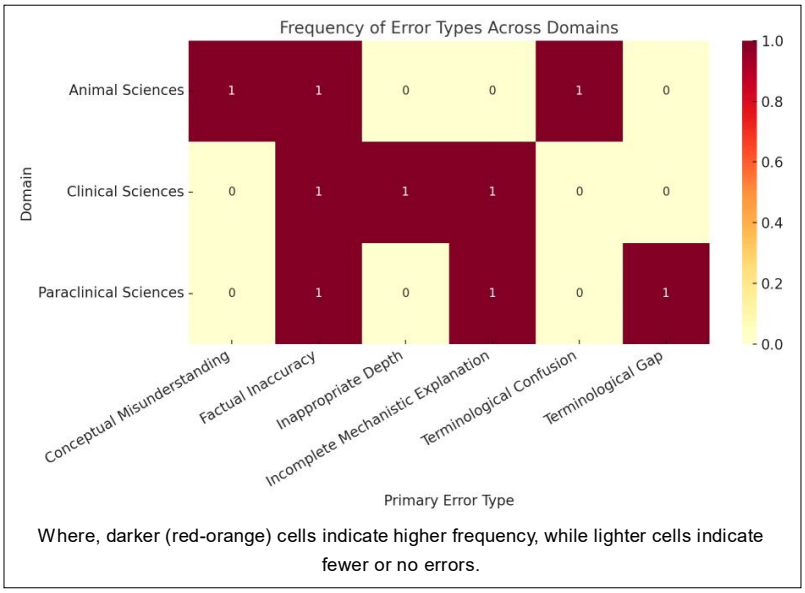

Fig 1: Color-coded heatmap showing the frequency of each error type across domains.

Table 5: Cohen's kappa (Inter-rater agreement).

Metric	Value	Interpretation
Cohen's kappa	-0.018	Poor agreement

- **Ambiguity in reasoning** (found in both models but less in ChatGPT-4.0).

Despite ChatGPT-4.0's stronger quantitative performance, these qualitative coding results highlight that some limitations remain, particularly in domain-specific terminology precision and consistent contextual alignment.

Our comparative evaluation of ChatGPT-3.5 and ChatGPT-4.0 across multiple veterinary science domains revealed statistically significant differences in performance, with ChatGPT-4.0 consistently achieving higher mean scores. One-way ANOVA analyses demonstrated that domain-specific score variations were significant, a finding reinforced by two-way ANOVA results indicating a strong model effect and a notable interaction between model type and academic domain. These outcomes suggest that ChatGPT-4.0 not only provides generally more accurate and comprehensive answers but also exhibits greater adaptability across disciplinary contexts. Similar trends have been documented in prior AI benchmarking studies in medical and veterinary education, where GPT-4 outperformed earlier iterations in factual accuracy, reasoning complexity and context-aware responses (Nori *et al.,* 2023; Wulcan *et al.,* 2025).

Despite these performance gains, the inter-rater reliability for flagged responses, as assessed by Cohen's Kappa, was extremely low (κ = -0.018), indicating poor agreement between evaluators. This aligns with earlier research in educational assessment demonstrating that subjective scoring of open-ended responses can result in substantial variability, particularly when raters weigh criteria such as factual correctness, clarity and domain relevance differently (McHugh, 2012). The negative Kappa value suggests that rater judgments were not only inconsistent but, in some cases, inversely correlated, reflecting the inherent subjectivity in qualitative evaluation of AI-generated answers.

The thematic coding of flagged responses revealed recurring patterns of model shortcomings. The most frequent themes included factual inaccuracies, lack of domain specificity and overgeneralization, with a smaller proportion of flags related to ambiguous phrasing and irrelevant content. This distribution mirrors patterns observed by Zhang *et al.* (2024) in AI-generated medical education content, where even high-performing models occasionally produced domain-inaccurate statements or resorted to generic phrasing when specialized knowledge was insufficiently represented in the training data.

Interestingly, the magnitude of improvement from GPT-3.5 to GPT-4.0 was most pronounced in Animal Sciences and Clinical Sciences, whereas Paraclinical Sciences showed a smaller relative gain. This may be attributable to uneven domain exposure during model training; prior studies have shown that large language models tend to perform better in areas with abundant, high-quality, publicly available literature compared to niche subfields with limited training representation (Wulcan *et al.,* 2025; Nori *et al.,* 2023).

In contrast to findings from controlled medical question-answering benchmarks that reported near-perfect rater agreement for structured, fact-based queries (Kung *et al.,* 2023), our results underscore the importance of robust scoring protocols and evaluator calibration when assessing AI performance in integrative, interdisciplinary domains like veterinary sciences. The low agreement in our study may be partly due to the complexity of veterinary problem-solving, which often requires synthesizing multi-domain knowledge, thus increasing the scope for interpretive variability among evaluators.

Table 6: Thematic coding matrix of flagged responses.

Domain	Q.No.	Question	ChatGPT-3.5 response (Score)	ChatGPT-4.0 response (Score)	Primary error type	Thematic code	Explanation of Issue
Animal sciences	4	Signs of heat in goats	Restless, mounting, swelling (7)	Mounting, discharge, vocal, tail wag (9)	Factual inaccuracy	Diagnostic omission	Missed key reproductive signs such as discharge and tail wagging.
Animal sciences	6	Define heritability	Trait passed to offspring (7)	% phenotypic variance due to genetics (10)	Conceptual misunder-standing	Overgenerali-zation	Confused genetic transmission with statistical measure of heritability.
Animal sciences	8	Qualitative vs quantitative traits	One gene vs many genes (7)	Inheritance patterns and examples (10)	Terminological confusion	Incomplete definition	Missed inheritance context and relevant examples.
Paraclinical sciences	4	Endotoxins vs exotoxins	Endotoxins = gram-neg.;	Exotoxins = proteins (7)	Endotoxins = lysis products;	Exotoxins = secreted (9)	Factual Inaccuracy Misclassification Over-simplified and taxonomically misleading categorization.
Paraclinical sciences	8	Adverse effects of NSAIDs	GI ulceration, kidney issues (7)	Gastric ulcers, nephrotoxicity, platelet inhibition (10)	Terminological gap	Non-technical Language	Used lay terms (“kidney issues”) instead of precise pharmacological terms.
Paraclinical sciences	10	Shock types	Circulatory failure;3 types listed (7)	Hypovolemic, cardiogenic, septic; mechanisms (10)	Incomplete Mechanistic Explanation	Oversimp-lification	Lackedpatho-physiological details for each shock type.
Clinical sciences	7	Symptoms of milk fever in cows	Weakness, low calcium (7)	Recumbency, cold ears, tremors, hypocalcemia (9)	Factual inaccuracy	Diagnostic omission	Missed hallmark signs such as cold ears and tremors.
Clinical sciences	8	Treatment of ketosis in dairy cows	Glucose (6)	IV glucose, propylene glycol, diet management (10)	Inappropriate depth	Partial treatment protocol	Provided only single therapeutic option without nutritional management.
Clinical sciences	9	Retained placenta causes	Nutrition, infections (6)	Hypocalcemia, uterine atony, infections (9)	Incomplete mechanistic explanation	Oversimplifi-cation	Omitted key physiological causes such as uterine atony.

Each entry in the table lists the question, scored responses from ChatGPT-3.5 and ChatGPT-4.0, the primary error type and its thematic code. Thematic Code Legend: Diagnostic Omission - Missing essential signs/symptoms for accurate identification; Overgeneralization - Overly simplified concept lacking statistical or mechanistic detail; Incomplete Definition - Partially correct but missing essential contextual elements; Misclassification - Incorrect or misleading categorization; Non-technical Language - Use of lay terms instead of discipline-specific terminology; Oversimplification - Lack of mechanistic or causal explanation; Partial Treatment Protocol – Missing comprehensive management strategies.

From a pedagogical perspective, our results support the integration of GPT-4 into veterinary education as a supplementary tool, particularly in areas where it demonstrated strong factual accuracy and contextual adaptation. However, the flagged content analysis highlights the need for human oversight and domain-specific fine-tuning before deploying such models in high-stakes educational or clinical contexts.

Collectively, the present findings both corroborate and extend prior AI evaluation literature: While GPT-4 represents a measurable advancement over GPT-3.5 in knowledge accuracy and adaptability, evaluator disagreement and domain-dependent performance gaps remain significant limitations. Addressing these issues will require a dual approach-algorithmic improvement in domain-specific reasoning and methodological refinement in evaluation frameworks.

CONCLUSION

In conclusion, our results underscore a clear generational leap in ChatGPT's capacity to deliver accurate, relevant and pedagogically valuable content in veterinary and animal sciences. The combination of statistical significance, high inter-rater reliability and thematically richer explanations in ChatGPT-4.0 positions it as a viable adjunct to traditional teaching modalities-though not yet a substitute for domain expertise and critical human judgement.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

AlZubi, A.A. and Al-Zu'bi, M. (2023). Application of artificial intelligence in monitoring of animal health and welfare. *Indian Journal of Animal Research.* **57(11):** 1550-1555. doi: 10.18805/IJAR.BF-1698.

Choudhary, O.P., Saini, J., Challana, A., Choudhary, O., Saini, J. and Challana, A. (2023). ChatGPT for veterinary anatomy education: An overview of the prospects and drawbacks. *Int J Morphol.* **41(4):** 1198-1202.

Kung, T.H. Cheatham, M., Medenilla, A., Sillos, C.J., Leon, L.D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J.J., Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health.* **2(2):** e0000198.

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V. and Eger, S. (2024). Chatgpt: A meta-analysis after 2.5 months. *Machine Learning with Applications*. **16:** 100541.

McHugh, M.L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica.* **22(3):** 276-282.

Nori, H., King, N., McKinney, S.M., Carignan, D. and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. (2024). [online]. Website https://openai.com/ [accessed 1 February, 2024].

Priya, D., Triveni, D., Singh, M., Gyanendra, G.K. and Vivek, S. (2024). Artificial intelligence: As an intelligent tool for the future animal production and better management: A review. *Agricultural Reviews.* **45(2):** 229-238. doi: 10.18805/ag.R-2297.

Reuters. (2023). ChatGPT sets record for fastest growing use base - analyst note [online]. Website https://economictimes.indiatimes.com/tech/technology/chatgpt-sets-record-for-fastest-growing-user-base-analyst note/articleshow/97542869.cms?from=mdr. [accessed 3 February, 2024].

Samala, A.D. and Amanda, M. (2023). Immersive learning experience design (ILXD): Augmented reality mobile application for placing and interacting with 3D learning objects in engineering education. *International Journal of Interactive Mobile Technologies.* **17(5):** 22-35. doi: 10.3991/ijim.v17i05. 37067.

Samala, A.D., Zhai, X., Aoki, K., Bojic, L. and Zikic, S. (2024). An In-depth review of ChatGPT's Pros and cons for learning and teaching in education. *International Journal of Interactive Mobile Technologies.* **18(2):** 96-117. doi: 10.3991/ijim.v18i02.46509

Van Dis, E.A., Bollen, J., Zuidema, W., Van Rooij, R. and Bockting, C.L. (2023). ChatGPT: Five priorities for research. *Nature*. **614(7947):** 224-226.

Wulcan, J.M., Jacques, K.L., Lee, M.A., Kovacs, S.L., Dausend, N., Prince, L.E. and Keller, S.M. (2025). Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. *Frontiers in Veterinary Science.* **11:** 1490030.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.

Zhang, C., Liu, S., Zhou, X., Zhou, S., Tian, Y., Wang, S. and Li, W. (2024). Examining the role of large language models in orthopedics: Systematic review. *Journal of Medical Internet Research.* **26:** e59607.