doi:10.3233/SHTI241092

Horse Diagnosis and Triage Accuracy of GPT-4o

Laura HAASE[a,b,1], Dagmar MONETT[b] and Martin SEDLMAYR[a]
[a] *Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany*
[b] *Department of Cooperative Studies – Computer Science, Berlin School of Economics and Law, Berlin, Germany*
ORCiD: Laura Haase https://orcid.org/0000-0003-1632-7334,
Dagmar Monett https://orcid.org/0000-0001-5750-972X,
Martin Sedlmayr https://orcid.org/0000-0002-9888-8460

Abstract. Animal owners may increasingly rely on large language models for gathering animal health information alongside internet sources in the future. This study therefore aims to provide initial results on the accuracy of ChatGPT-4o in triage and tentative diagnostics, using horses as a case study. Ten test vignettes were used to prompt situation assessments from the tool, which were then compared to original assessments made by a veterinary specialist for horses. The most probable diagnosis suggested by ChatGPT-4o was found to be quite accurate in most cases, with the urgency to contact a veterinarian sometimes assessed as higher than necessary. When provided with all relevant information, the tool does not seem to compromise horse health by recommending excessively long waiting times, although there is still potential for improving the relief of veterinarians' workload.

Keywords. Artificial Intelligence, LLM, Animals, Horses, Diagnosis, Triage

1. Introduction

Animal owners search the Internet for information regarding their pet's health [1], partly due to the substantial workload of veterinarians in Germany and elsewhere, which jeopardizes emergency services [2]. However, online animal health information quality varies, e.g. is sometimes inaccurate or incomplete [1].

Recently, large language models (LLMs), particularly GPT, gained popularity for the use in medical diagnostics/ triage, demonstrating promising first results [3,4]. ChatGPT, a dialogue-based artificial intelligence chatbot, is trained on extensive datasets using a transformer architecture that allows the underlying neural network to process large chunks of input at a time, enabling it to predict contextually plausible combinations of words and sentences as an answer to a given prompt [5]. However, this technology comes with several limitations including; limited reasoning and focusing capabilities, hallucinating or reproducing inaccurate information, being highly sensitive to changes in prompts, and ignoring human fallibility [5–7]. It is therefore important to carefully evaluate LLM use in critical contexts like the medical field [5].

[1] Corresponding Author: Laura Haase, Department of Cooperative Studies – Computer Science, Berlin School of Economics and Law, Alt-Friedrichsfelde 60, 10315 Berlin, DE; E-mail: laura.haase@hwr-berlin.de.

Currently, research on the accuracy of LLMs in veterinary medicine is scarce, primarily focusing on educational assessments or specific pathologies [7,8]. Thus, this study aims to provide an exploratory assessment of the accuracy of ChatGPT-4o in the area of veterinary triage (to support veterinarians) and general preliminary diagnostics (to meet the information needs of animal owners [9,10]). For this purpose, a case study involving horse medicine was chosen.

2. Methods

Ten test vignettes developed with a veterinary specialist for horse health [11] were used for this study, each including the age, gender and breed of a horse, as well as the symptoms that the horse owner may observe for a present condition. Additionally, vignettes provide the correct diagnosis and an assessment of the urgency for a veterinary appointment. A situation assessment for each test case was created using ChatGPT-4o in the German language, most recently on June 17, 2024. The tool was instructed to provide an assessment of the urgency for contacting a veterinarian and the most likely diagnoses from the perspective of a veterinary specialist for horses. The prompt specified that if multiple diagnoses were possible, they should be ordered by descending probability. Each case assessment was prompted independently.

Tentative diagnoses provided by the tool were evaluated according to the following categories: *(A) most probable diagnosis (ChatGPT) matches the correct diagnosis in the vignette*, *(B) most probable diagnosis (ChatGPT) is closely related to the correct diagnosis, but contains some inaccuracies*, *(C) correct diagnosis in the vignette is not given as most probable, but included among the most probable diagnoses (ChatGPT)*, and *(D) correct diagnosis in the vignette is not among suggested diagnoses (ChatGPT)*.

The urgency assessments provided by the tool were also categorized into groups: *(a) urgency assessment (ChatGPT) matches urgency assessment in the vignette*, *(b) urgency assessment (ChatGPT) is higher than the urgency assessment in the vignette*, and *(c) urgency assessment (ChatGPT) is lower than the urgency assessment in the vignette*.

3. Results

ChatGPT-4o generated four to six tentative diagnoses for each vignette. Each diagnosis was accompanied by a brief description of the condition, often supplemented by possible associated symptoms. Additionally, each output from the tool provided a brief explanation of the urgency assessment and, in some cases, included further action recommendations. A summarized overview of the generated tentative diagnoses and urgency assessments, as well as the correct information from the test vignettes and their categorization, is presented in Table 1.[2] The correct diagnosis has been proposed by the tool in 90% of the cases (80% as most probable diagnosis), with the remaining 10% being closely related to the correct diagnosis. The urgency was accurately assessed in 70% of the cases, while in the persisting 30%, the assessments were classified with a higher degree of urgency compared to the veterinarian's evaluation.

[2] Original prompts and answers generated by the tool, as well as their translation into English language can be found in Supplementary File 1: https://cloud.hwr-berlin.de/s/4nrqSRE6aoLTmyq (pw: EFMI_STC_24).

Table 1. Results of the vignette tests using ChatGPT-4o.

Diagnosis (vignette)	Diagnosis (ChatGPT)	Evaluation	Urgency (vignette)	Urgency (ChatGPT)	Evaluation
Laminitis	(1) Laminitis (2) Tetanus (3) Botulism (4) Myositis (5) Infectious diseases	*A*	**vet visit on the same day**	Immediate contact, even in emergency services	*a*
Colic	(1) Gas colic (2) Constipation colic (3) Spasmodic colic (4) Sand colic (5) Intestinal twisting/ displacement	*A*	**vet visit as soon as possible**	Immediate contact, even in emergency services	*a*
Conjunc-tivitis	(1) Conjunctivitis (2) Allergic reaction (3) Corneal injury/ ulceration (4) Uevitis (5) Infectious keratitis	*A*	**vet visit no later than the following day**	Prompt contact, at the latest the next day during regular opening hours (examination within 24 hours)	*a*
Narcolepsy	(1) Narcolepsy (2) Equine Protozoal Myeloencephalitis (3) Equine Metabolic Syndrome/ Cushing's Syndrome (4) Equine Sarcoids/ other tumors in the brain/ spinal cord (5) Myasthenia gravis	*A*	**vet visit within 1-2 weeks**	Immediate contact, even in emergency services	*b*
Heat stroke	(1) Equine influenza (2) EHV-1 (3) Pneumonia (4) Colic (5) Tetanus (6) Heat stroke	*C*	**vet visit on the same day/ as soon as possible**	Immediate contact, even in emergency services	*a*
Gastritis	(1) Gastric ulcers (2) Dental problems (3) Intestinal parasites (worm infestation) (4) Reflux disease (5) Liver/ kidney problems	*B*	**vet visit within a week**	Immediate contact, even in emergency services	*b*
Esophageal obstruction	(1) Esophageal obstruction (2) Aspiration pneumonia (3) Dysphagia (4) Colic (5) Equine influenza/ other respiratory disease	*A*	**vet visit as soon as possible**	Immediate contact, even in emergency services	*a*
Equine asthma	(1) Chronic obstructive bronchitis/ equine asthma (2) Infectious respiratory disease (3) Allergic reactions (4) Heart problems (5) Parasitic infestation (6) Atypical pneumonia	*A*	**vet visit within a week**	Immediate contact, even in emergency services	*b*

EHV-1	(1) EHV-1 (2) Influenza (3) Strangles (4) Equine viral arteritis (5) Fungal infection/ bacterial meningitis/encephalitis (6) West Nile virus	*A*	**vet visit as soon as possible**	Immediate contact, even in emergency services	*a*
Picked-up nail	(1) Picked-up nail/ hoof abscess (2) Nail puncture injury (3) Pododermatitis (4) Bruise of the hoof corium	*A*	**vet visit no later than the following day**	Immediate contact, even in the emergency service (exception: with mild symptoms and no severe pain shown, horse can be monitored and contact can be delayed until the next day during regular opening hours; examination within 24 hours)	*a*

4. Discussion and Conclusions

Except for one test case, the most likely diagnosis generated by ChatGPT-4o matched the correct diagnosis from the vignette, although sometimes it was not named completely accurate (gastritis and gastric ulcers are normally treated separately in the literature but both describe similar inflammatory and ulcerative changes in the gastric mucosa and can represent different stages of the same illness [12]) or the output was using outdated/ imprecise terminology (equine asthma is a new term for conditions previously differentiated into different conditions like chronic obstructive bronchitis and inflammatory airway disease [13]). Additionally, the level of detail in the responses varied; e.g. while most tentative diagnoses used the term colic generically, one case distinguished between different types of colic. With those results, horse owners would roughly be guided towards appropriate further research topics and considerations. Nevertheless, to prevent potential risks for the animal's health, owners should be educated about limitations of the tool and the importance of professional advice [10].

The tools urgency assessment for contacting a veterinarian was either consistent with or higher than the veterinarian's assessment. This indicates that the horses' health would unlikely be compromised by the tool's recommendations to delay veterinary consultation. However, a more precise assessment would be desirable to reduce the veterinarians' workload, particularly during emergency service hours. Higher urgency assessments were consistently associated with vignettes featuring symptoms of prolonged duration. As this reduces the probability of a sudden, significant deterioration, those vignettes received a lower urgency assessment in the process of vignette creation.

The results generated by ChatGPT-4o in this study demonstrated performance in horse medicine comparable to that reported for GPT-3 in human medicine [3]. Similar findings have been observed in human medical triage decisions involving GPT-4 [4].

The conducted exploratory study has several limitations that must be acknowledged. The sample size of test cases is small, and each prompt was only executed once. Additionally, prompts were designed by computer scientists, who may have used more precise wording than typical users, to which LLMs react sensitively [6]. Furthermore, it is important to consider that the symptom sets provided by animal owners may differ from those in the test vignettes created by veterinarians [9] and the tool does not assess

accuracy or completeness of the input data [5]. Consequently, the accuracy of the produced results might not be generalizable.

Future work should involve expanding the number of test vignettes and repeating result generation to enhance the robustness of findings. Additionally, a more thorough examination of the ethical implications of LLM usage by animal owners is recommended. Other studies report superior performance of LLMs like Claude 3 Opus, when compared to various GPT versions [14]. Accordingly, comparative analyses among multiple tools should be conducted within this context as well. Regardless the algorithm/ tool used, supporting animal owners in the process of gathering/ applying information concerning their animals' current health situation is advisable [5,9,10].

References

[1] Kogan L, Oxley JA, Hellyer P, Schoenfeld R, Rishniw M. UK pet owners' use of the internet for online pet health information. Veterinary Record. 2018 May;182(21):601–601, doi: 10.1136/vr.104716.

[2] Bundesverband Praktizierender Tierärzte e. V. Tierärztemangel: Tiergesundheit und Tierschutz in akuter Gefahr [Internet]. bpt - Bundesverband Praktizierender Tierärzte e.V. 2022 [cited 2022 Jan 9]. Available from: https://www.tieraerzteverband.de/bpt/presseservice/meldungen/2022/2022_11_14_EuroTierpk.php

[3] Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, Beam A. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model [Internet]. 2023 [cited 2024 Jun 11]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.01.30.23285067

[4] Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT. The American Journal of Emergency Medicine. 2024 Apr;78:170–5, doi: 10.1016/j.ajem.2024.01.037.

[5] Burtsev M, Reeves M, Job A. The Working Limitations of Large Language Models. MIT SMR. 2023 Nov 30;(Winter 2024):2–4.

[6] Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, Vielhauer J, Makowski M, Braren R, Kaissis G, Rueckert D. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024 Jul 4;1–23, doi: 10.1038/s41591-024-03097-1.

[7] Chu CP. ChatGPT in veterinary medicine: a practical guidance of generative artificial intelligence in clinics, education, and research. Front Vet Sci. 2024 Jun 7;11:1395934, doi: 10.3389/fvets.2024.1395934.

[8] Abani S, De Decker S, Tipold A, Nessler JN, Volk HA. Can ChatGPT diagnose my collapsing dog? Front Vet Sci. 2023 Oct 10;10:1245168, doi: 10.3389/fvets.2023.1245168.

[9] Haase L. Analysis of the Usage Context of an mHealth Application for Equestrians. In: Röhrig R, Grabe N, Haag M, Hübner U, Sax U, Oliver Schmidt C, Sedlmayr M, Zapf A, editors. Proceedings of the 68th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology eV (gmds); 2023 Sep 17-21; Heilbronn, Germany. Amsterdam: IOS Press; 2023. p. 117–125, doi: 10.3233/SHTI230702.

[10] Jokar M, Abdous A, Rahmanian V. AI chatbots in pet health care: Opportunities and challenges for owners. Veterinary Medicine & Sci. 2024 May;10(3):e1464, doi: 10.1002/vms3.1464.

[11] Haase L, Rahn L. Konzeptionierung einer Softwareanwendung zur Verdachtsdiagnostik und Triage in der Pferdemedizin. In: Eggert S, Lemke C, Majuntke V, Malzahn B, Meister VG, Simbeck K, Czarnecki C, Wolf M, editors. Angewandte Forschung in der Wirtschaftsinformatik 2022: Tagungsband zur 35. Jahrestagung des Arbeitskreises Wirtschaftsinformatik an Hochschulen für Angewandte Wissenschaften im deutschsprachigen Raum (AKWI); 2022 Sep 11-13; Berlin, Germany. Berlin: GITO mbH Verlag; 2022. p. 232–44, doi: 10.30844/AKWI_2022_15.

[12] Brehm W, Gehlen H, Ohnesorge B, Wehrend A, Dietz O, Huskamp B, Bartmann CP, editors. Handbuch Pferdepraxis. 4., vollständig überarbeitete und erweiterte Auflage. Stuttgart: Enke Verlag; 2017. 1239 p.

[13] Pferdeklinik Aschheim. Husten bei Pferden: Akute und chronische Atemwegserkrankungen [Internet]. Pferdeklinik Aschheim. [cited 2024 Jun 11]. Available from: https://www.pferdeklinik-aschheim.de/husten-bei-pferden/

[14] Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, Gonoi W, Abe O. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol. 2024 Jul, doi: 10.1007/s11604-024-01619-y.