

ChatGPT versus veterinary students for clinical data extraction: a mixed-methods study

Yumi Chang

Texas A&M University

Armanto Sutedjo

Texas A&M University

Candice P. Chu

cchu@cvm.tamu.edu

Texas A&M University

Article

Keywords: Artificial Intelligence, Data Mining, Dogs, Electronic Health Records, Large Language Models, Natural Language Processing

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10758356/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Most scientific research relies on structured data, yet veterinary medical records are predominantly free text, making evidence-based clinical research difficult to scale. Data extraction is therefore necessary, but manual extraction by domain experts is slow and costly; while artificial intelligence (AI) offers a scalable alternative, its performance relative to humans remains poorly understood. This study involved 10 second-year veterinary students extracting data from medical records of three suspected canine pancreatitis cases, first manually, and then with ChatGPT. We compared extraction accuracy and time, and assessed student perceptions through pre- and post-surveys followed by semi-structured interviews. ChatGPT was 34 times faster than student extractors (15 seconds versus 8.5 minutes per record, $p < 0.01$), and both methods achieved $> 90\%$ accuracy. Unexpectedly, over-inference happened more in humans than AI, reflecting the prompt's role in conservative extraction. Students generally preferred AI-assisted extraction over manual extraction, and deductive thematic analysis identified six themes, including AI's ability to offload repetition and reduce the cognitive burden of text interpretation. AI can efficiently and accurately automate veterinary data extraction, but human judgment remains essential for complex narratives.

Introduction

Animal health data proves valuable in various research contexts, including disease surveillance^{1,2}, epidemiological studies³, antimicrobial usage pattern analysis⁴, and specific disease predictions⁵⁻⁷. All these impactful studies require large amounts of health-related data (e.g., signalments, vital signs, and diagnostic results) that are commonly stored in free-text form in electronic health records (EHRs). Before these data can be analyzed, they must be extracted and converted into a structured, machine-readable format. However, because free-text data are 'messy'⁸, data extractors (e.g., student workers) must be familiar with medical terminology and sufficiently experienced to form a complete clinical picture from notes written by practitioners⁹. Furthermore, even when domain expertise is available, manual data extraction from large amounts of EHRs remains labor-intensive, time-consuming, and error-prone, ultimately hindering clinical research⁹.

Automated data extraction using artificial intelligence (AI) provides a potential solution to these constraints¹⁰. The most common approach stems from the subfield of natural language processing, in particular because of powerful advancements in the application of large language models (LLMs). One trend in the use of LLMs is to structure research-specific data from EHRs in both human and veterinary medicine¹⁰. Because extracted data serves as an intermediate product, errors during extraction can propagate into downstream research, making assessment of AI-curated data quality a critical safeguard^{11,12}.

Evaluating the performance of AI-assisted data extraction centers primarily on accuracy, with human expert consensus serving as the ground truth across both veterinary^{5,13,14} and human medicine¹⁵⁻¹⁷.

Time efficiency is an equally important but less frequently evaluated performance metric. AI-assisted data extraction is intuitively expected to be faster than manual review, but there is limited empirical evidence—prior studies only provided rough estimates of extraction time^{5,15}. For researchers integrating LLMs into data extraction workflows, the critical question is whether time efficiency comes at the cost of compromising data quality relative to traditional manual extraction. To our best knowledge, no study has directly compared manual and AI-assisted data extraction with respect to both accuracy and time. This study fills the gap by performing objective measurements of both time and accuracy of LLM-assisted EHR extraction.

Importantly, while performance evaluation reveals the power of AI, implementing these tools in clinical research also depends on workforce acceptance. For example, undergraduate and veterinary students often assist with data extraction and other repetitive research tasks and, as future practitioners and researchers, are likely to become important users of AI-assisted tools in both clinical and research settings. Previous studies examined veterinary students' attitudes toward AI, self-perceived AI literacy, and usage frequency^{18–21}, but none has evaluated students' perceptions after completing hands-on AI tasks. Moreover, those studies relied primarily on surveys, providing limited insight into the reasoning and experiences underlying user responses. Thus, our mixed-methods study incorporated both surveys and interviews with veterinary students performing AI-assisted data extraction, using the Technology Acceptance Model (TAM), which explains individuals' intentions to adopt a technology based primarily on its perceived usefulness and perceived ease of use^{22,23}.

This study is the first to simultaneously quantify the time, accuracy, and perception of AI-assisted clinical data extraction in veterinary students. Specifically, we aimed to: (1) compare time spent between manual and AI-assisted extraction of clinical data from veterinary electronic health records, (2) compare the accuracy and characterize the errors made in both methods, and (3) examine students' perceptions of and attitudes toward AI-assisted extraction through TAM-based surveys and semi-structured interviews.

Methods

Human Subjects

Eligibility was limited to second-year veterinary students enrolled in an elective course “*Artificial Intelligence and Digital Tools for Next-Generation Veterinarians*” (AI literacy course in Fig. 1) in the College of Veterinary Medicine & Biomedical Science at Texas A&M University²⁴, and all 10 enrolled students participated in the study. The course, developed and taught by the corresponding author (CPC), provided relevant background knowledge regarding the fundamentals of AI and LLMs. All activities were required as homework for the course, but participation in the study was entirely voluntary and did not affect course grading. All participants read and signed the written informed consent before enrollment. The Texas A&M University Institutional Review Board approved the comparative extraction task (STUDY2025-0206) and the survey and interview protocols (STUDY2025-1493) before the study began. All methods were performed in accordance with relevant institutional guidelines and regulations.

Overall Study Design

Quantitative and qualitative components were structured into a sequential, integrated workflow (Fig. 1) to capture how these two components mutually inform one another. First, a pre-extraction survey was conducted in the first week of the AI literacy course to assess students' prior experience with data extraction and LLMs. Then, for the data extraction phase, students performed manual extraction, completed a one-month washout period, and lastly performed AI-assisted extraction. The final phase comprised a post-extraction survey and a follow-up semi-structured interview.

Data Extraction

Manual and AI-assisted extraction periods (Fig. 1) each spanned 2 weeks. In the first week, participants received one practice case to familiarize themselves with the procedure. All returned submissions were verified to ensure they adhered to the protocol, and the results from the practice case were not included in the study. In the second week, participants received three additional study cases and were allowed to complete each case at their own pace during the two-week period. They were instructed to record the start and finish timestamps (24-hour format, HH:MM:SS) for both methods. To minimize carryover effects from the prior task, a washout period of approximately one month separated the two tasks, with all participants completing the manual extraction first before proceeding to the AI-assisted task.

Medical records

Four synthesized canine cases (one practice case and three actual cases) with suspected acute pancreatitis generated from patient records archived in the Texas A&M University Veterinary Medical Teaching Hospital's electronic medical record system were used for both manual and AI data extraction. Each case was compiled as a single PDF (ranging 3 to 4 pages), including clinical history, physical examination findings, laboratory values (complete blood count, serum chemistry, and canine pancreatic lipase), and ultrasound reports. We used a pre-established extraction template (Excel file) comprising 63 variables that included both categorical and numerical data¹⁰ (Table 1). **Supplementary Information S1** provides all study information supplied to participants, including the practice case, the Excel data entry template, and instructions detailing the extraction procedures.

Table 1

Summary of predefined data items used for extraction. Items were grouped by record section and type, comprising a total of 63 variables. Categorical variables were coded using predefined categories; numerical variables required entry of values. Representative examples are provided for each record section; for the complete list of items and their categories/units, see the extraction template in **Supplementary Information S1**.

History	Categorical items	Numerical items	Total
	n = 10 • Gender • Neuter status • Presence of certain clinical signs (vomit, diarrhea, weight loss...)	n = 1 • Age	11
Physical examination	n = 3 • Hydration status • Abdominal palpation • Rectal exam	n = 5 • Body weight • Body temperature • Heart rate • Respiratory rate • Body condition score	8
Complete blood cell	n = 0	n = 15 • White blood cell • Reb blood cell count • Hemoglobin...	15
Chemistry	n = 0	n = 21 • Glucose • Lactic acid • Creatinine...	21
Ultrasound report	n = 8 • Pancreatic size • Pancreatic echogenicity • Pancreatic echotexture...	n = 0	8
Total	21	42	63

AI tool and prompts

The AI-assisted extraction was conducted using ChatGPT (GPT-5.2, an LLM) via the TAMU AI Chat (<https://tamu.ai/>), a university-approved platform aligned with data security compliance and accessible to all Texas A&M University members. The structured prompt for AI-assisted extraction is provided in the **Supplementary Information S1**.

Data and statistical analysis

Extraction time was calculated by subtracting the start timestamp from the finish timestamp for each method. Accuracy was computed as the number of correct extractions divided by the total number of extractable data elements ($n = 63$). The ground truth was established in an unpublished study by two PhD students with a DVM background who independently extracted the same cases, with consultation with an internist for any disagreement. Normality of the time and accuracy data was assessed using Shapiro-Wilk test to determine the appropriate paired statistical test for comparing extraction methods. Statistical significance was defined as $p < 0.01$.

Perceptions of AI-Assisted Extraction

Perceptual analysis was conducted using a mixed-methods approach through surveys and interviews.

Surveys

Surveys were distributed via Qualtrics at two timepoints: at the beginning of the AI literacy course (Week 1, before the practice manual data extraction) and after the final AI-assisted data extraction (Week 10) (Fig. 1). The pre-extraction survey consisted of eight questions that gathered baseline data on participants' prior experience with manual information extraction ($n = 3$), LLM usage ($n = 2$), self-reported technological proficiency ($n = 1$), initial trust in LLMs ($n = 1$), and native language ($n = 1$), using Likert, binary, ordinal, frequency, and multiple-choice response formats. The post-extraction survey questions were derived from the TAM^{22,23}. The post-extraction survey included 12 questions: perceived usefulness ($n = 5$), perceived ease of use ($n = 2$), behavioral intention to use ($n = 2$), users' perceived liability of usage ($n = 1$), and attitudes toward AI in veterinary education ($n = 2$). Ten items used a 5-point Likert scale, and two were open-ended. The complete questionnaires are provided in **Supplementary Information S2**.

Interviews

Semi-structured interviews were conducted to gain deeper insights not captured by the survey questions. Interview questions were centered around participants' survey responses to probe their attitudes. Interview guidelines are provided in **Supplementary Information S2**. Among 10 participants, seven completed face-to-face interviews in distraction-free rooms on campus, and three completed interviews virtually via Zoom. Interviews lasted between 10 and 30 minutes and were conducted by the same interviewer (YC). With participants' consent, all interviews were audio-recorded and transcribed

using the built-in transcription feature of Apple Voice Memos on an iPhone 13 (iOS 18.4, Apple, CA, USA). Transcripts were reviewed and corrected by the interviewer against the original recordings. Finalized transcripts were paired with survey responses and compiled as PDFs for qualitative analysis. Given the TAM structure embedded in this study, all authors coded the transcripts independently and deductively to identify common concepts aligned with the framework. As unexpected ideas emerged beyond the predefined topics, a hybrid deductive–inductive approach was adopted to form a comprehensive thematic picture. One author (YC) merged the three independent preliminary codebooks, and the authors established the final code labels through discussion and consensus.

Results

Manual Versus AI-assisted Data Extraction

Data were generated from 10 students who each completed extraction across three actual cases using both manual and AI-assisted methods, for a total of 30 paired extractions. Data cleaning revealed that one student's AI-assisted responses for Cases 2 and 3 were identical to those of Case 1, likely due to a copy–paste or submission error. Since this was considered participant non-compliance with the task protocol rather than a performance issue of the AI tool, data from this participant were excluded from the following analyses, resulting in a final sample of nine students for time and accuracy comparisons.

Extraction time across all three cases was compared between methods (Fig. 2). The median total extraction time for AI-assisted extraction was 46 seconds, compared to 1,572 seconds (approximately 26 minutes) for manual extraction. Stratified by case, manual extraction time demonstrated a sequential decline across cases (mean/ median: Case 1–575.4/603s; Case 2–518.3/523s; Case 3–504.3/415s), averaging 513.6s (around 8.6 min); AI-assisted extraction time remained consistent across all three cases (mean/ median: Case 1–28.6/14s, Case 2–23.3/18s, Case 3–19.8/13s), averaging 15s. Normality tests confirmed normal distribution for both conditions, and a subsequent paired t-test revealed that AI-assisted extraction was significantly faster than manual extraction ($p < 0.01$).

Accuracy was calculated per student per case based on the number of correctly extracted variables out of 63 predefined data elements, yielding nine paired accuracy scores per method across three cases. The distribution of per-case and total accuracy for both manual and AI-assisted extraction across nine students is presented in Fig. 3.

For manual extraction, per-student mean accuracy ranged from 92.6% to 97.4%, with a median of 95.8%. For AI-assisted extraction, per-student mean accuracy ranged from 68.8% to 100%, with a median of 99.5%. The 68.8% value is an outlier caused by a systematic extraction error in Case 3 for one participant, in which the AI-generated output did not comply with the structured prompt, returning free-text descriptions rather than the predefined numeric codes specified in the prompt. For example, for chronicity of pancreatitis in ultrasound reports, the expected output was a numeric code (0: not specified; 1: acute; 2: acute-on-chronic; 3: chronic), but the AI returned the phrase 'acute on chronic

pancreatitis' rather than the code '2'. Even for numerical variables, values were returned with accompanying labels rather than as standalone numbers, for example, 'Red Blood Cell Count 3.75' rather than '3.75'. Although semantically correct, these non-compliant outputs were classified as invalid extractions and scored as errors for accuracy calculation because they required additional post-processing to be research-ready.

Due to the non-normal distribution of AI-assisted accuracy scores, a Wilcoxon signed-rank test was performed, revealing no statistically significant difference in mean accuracy between manual and AI-assisted extraction ($p = 0.125$).

To further characterize extraction errors, error counts were aggregated across all three cases to identify the variables with the most mistakes. Previously discussed submission errors and AI formatting issues in Case 3 for one participant were excluded, as they would obscure the true patterns of content-based errors (Fig. 4). For both manual and AI-assisted extraction, errors were concentrated in categorical variables (e.g., ultrasound reports and physical examination) rather than numerical values. Within the numerical variables, manual extraction errors ($n = 11$) were scattered across items, whereas the AI-assisted method was almost perfect, with only one error. Manual numerical errors were predominantly data-entry related. These included single-digit typos (e.g., 5.7 instead of 5.8) and row-shifting errors, in which a single misalignment propagated into three consecutive incorrect values. In contrast, one single AI numerical error occurred in Case 1 where the plasma protein value appeared as ** in the record, and the AI returned an empty cell rather than the ** symbol.

For categorical variables in the history section, manual errors came from clinical signs coded “absent” when present, or vice versa. These errors likely reflect that student extractors have difficulty extracting information from long, narrative case histories with substantial irrelevant detail. The AI-assisted method handled these narrative histories without any error. In the physical examination section, hydration status was another recurrent source of disagreement. This variable was not directly stated in Cases 2 and 3, hence the coding of “not mentioned” in the ground truth. However, four manual errors and one AI-assisted error were observed, implying that hydration status was inferred from related findings, such as the moisture of the mucous membrane.

Both humans and AI made errors extracting ultrasound reports, largely by misclassifying mentioned variables as absent or vice versa. A pattern unique to manual extraction was from the abdominal fluid effusion variable. Of the 11 errors for this variable, six involved ambiguous descriptors. The coding options were 0 (not mentioned), 1 (absence), 2 (fluid within normal limits), and 3 (presence). The reports used qualitative terms such as 'scant' or 'small volume,' which the ground truth classified as presence (3). Students, however, appeared to interpret these descriptors as fluid within normal limits (2). This error was less common in AI, suggesting a more conservative approach compared to that of human participants.

Perceptions of AI-Assisted Data Extraction

Quantitative Findings: Pre-Extraction Survey

All 10 enrolled students completed the survey assessing their previous extraction and LLM experience (Fig. 5). The cohort comprised four males (40%) and six females (60%), including nine native English speakers (90%) and one non-native speaker (10%). Most students (80%) reported being comfortable extracting information from PDF documents (Question 2, Q2), and more than half (60%) had prior experience extracting data from medical records (Q3). Among those with prior extraction experience ($n = 6$), one student (17%) had processed greater than 100 records, whereas the remaining five (83%) had processed less than 50 (Q4).

Sixty percent of the students perceived themselves as tech-savvy, with 30% being neutral and 10% expressing less savviness (Q5). Regarding LLM use, 70% of students reported prior experience with these tools (Q6). Among LLM users ($n = 7$), all had used ChatGPT, 57% had used Gemini, and one student had used Copilot (Q7).

Trust in LLMs varied along a gradient, ranging from “A little” (20%) and “A moderate amount” (50%) to “A lot” (20%) and “A great deal” (10%), with no students selecting “Not at all” before the extraction tasks (Q8).

Quantitative Findings: Post-Extraction Survey

Under the TAM framework, mental workload can serve as a proxy for perceived ease of use²⁵. After experiencing both manual and AI-assisted extraction, 70% of students agreed that manual extraction was mentally demanding (Q1), whereas 80% disagreed that AI-assisted extraction was mentally demanding (Q2) (Fig. 6). The intensity of response was more prominent for AI, with 70% strongly disagreeing that it was demanding, versus a more moderate distribution (30% strongly agree) for manual extraction.

In terms of perceived usefulness, most participants (70%) felt that AI-assisted extraction enhanced their case analysis (Q4), while 80% disagreed with the concern that AI would compromise their capacity for independent clinical judgment (Q5). These favorable perceptions carried over to perceived accuracy (Q7) and efficiency (Q9), both getting an overall 90% agreement. However, they differed in the strength of agreement. While 80% of participants strongly agreed that AI saved them time, only 20% expressed the same level of strong agreement that the AI-generated output was correct and complete, reflecting greater trust in efficiency than in accuracy.

Outside the TAM lens, 90% of respondents took responsibility for the AI-extracted data, with only 10% remaining neutral (Q8). Overall, all students preferred AI involvement over manual methods (Q3), though confidence in AI-assisted extraction varied (Q6).

Beyond the extraction task, 90% of students agreed that AI should play a larger role in veterinary education (Q10).

Qualitative Findings: Interviews

Six main themes were synthesized from the interview data (Table 2), including perceived usefulness, perceived ease of use, and behavioral intention to use, prior experience, perceived liability, and AI in veterinary education. Although presented separately, these themes are closely intertwined (Fig. 7). Because extraction was conducted digitally rather than on paper copies, individual technology preferences and proficiency likely shaped perceived ease of use with the AI workflow. This perception could then extend to perceived efficiency, as one participant explained:

It was faster once I figured it out. But like to get where, it was a little complicated to I was like, oh, maybe I would just rather do it myself.

The perceived usefulness of AI was not only about saving time, but also about reducing the mental effort involved in extraction. Participants described the manual process as repetitive and attention-demanding:

Mainly just very tedious. Going back and forth and reading and everything.

Beyond this repetition, the manual process required significant interpretive effort to extract meaning from complex clinical narratives, presenting a cognitive burden that AI was perceived to minimize:

That's definitely time-consuming and requires a little more out of you, like knowledge-wise, of like words, vocabulary, [...] but you need to kind of know how to perform it if you want to interpret somebody else's results. You want to kind of put yourself in their shoes and try to see what they saw.

Depending on how the medical record was written. Sometimes it's unclear, vague on a lot of things, and you have to, I guess, make that decision, and I think I would say maybe that's the messily texting part of it. While an AI model would kind of just decide for you.

Once AI was integrated into the workflow, perceived accuracy became part of a reciprocal loop involving perceived liability. Students' initial impressions were rooted in their trust in and exposure to LLMs, though some also acknowledged the risks of over-reliance:

I could miss some things if I'm like, AI hasn't failed me yet, then I just submit whatever record I extracted, and then I miss like the value or something.

Consequently, participants predominantly attributed responsibility to the human operator, which in turn motivated verification behaviors following AI extraction to secure their perceived accuracy. This verification, however, cycled back to influence their perception of overall efficiency, as one participant noted:

Why are you using it in the first place if you're going to double-check it anyway?

The relevance of these tools extended beyond the extraction task, with participants' attitudes shaped by workforce expectations and a hope for greater AI integration in education:

I knew that this is something that's going to become a huge part of our lives if it's not already. So I need to learn now and I need to kind of familiarize myself of how it works and how to use it [...], because this is kind of what every generation does.

Table 2

Summary of qualitative themes from student interviews on AI-assisted clinical data extraction. AI, artificial intelligence. PI, principal investigator.

Themes	Subthemes	Representative quotes
Prior experience	Prior extraction	"I'm a medical record reviewer, which basically just means I copy paste, do that every day, and put it into their database."
	Prior exposure to LLMs	"I use it sometimes to, like, check my documents just for like spelling you and grammar errors."
	Trust in LLMs	"I remember one of my classmates that has used ChatGPT first year compared to now, second year, and he said, like, it's way smarter... So I trust it a little. I do."
	Perceived technological proficiency	"I'm as tech savvy as my generation requires me to be." I'm just a vet student. Like, I use technology for, like, school, like the bare minimum." "I would just say, like, general basic knowledge to use any kind of electronics, so like computers, phones, and then, just like, basic knowledge"
	Technology preference	"I am a religious MacBook user. So I like never downloaded Excel."
Perceived usefulness	Perceived cognitive offloading	"It was most helpful to not only have the data extracted but also formatted to be pasted into the document instead of typing line by line in the results column."
		"You're distracted and then coming back to the task and then you kind of sometimes have to start over and you find your place instead of just picking up where you left off sometimes. Or even inputting. Like on the column, did I do the row that's here or here?"
		"Sometimes it comes down to like one number, you know? That you might miss just from your eyes."
		"I feel like AI should be for is like completing menial tasks that I shouldn't waste my time doing, like copying down information."
	Perceived functional advantages	"In like, clinical data extraction where you want it black and white, I think that it is, better than a human using their like knowledge to over interpret results."
	Perceived accuracy	"I saw errors, I think, in both. I think I only saw one or two in the AI and then with my own personal self. I don't remember. I'm not going to throw myself under the bus, but I did see myself make mistakes, for sure. " "I didn't, like, look at every nitty gritty, but it was just kind of like, um. just based off of like skimming through it, [...], it seemed pretty accurate. "
	Perceived efficiency	"I mean, it was a lot quicker to use AI for all parts."

Themes	Subthemes	Representative quotes
		"It generated that in less than a second, and I doubt it would take me more than 10 minutes to double check it." "I don't know how much time it would save with, like, a really large number of records to then still go through in fact check them all. [To] be honest. I don't know. Could end up being some amount of time." "I probably won't check it to the level I did whenever I had to read the whole thing."
Perceived ease of use	Difficulty	"I struggled a lot with the TAMU AI chatbot [for] whatever reason."
	Setup	"Because I used a double screen, I usually use a big screen and then my laptop's attached. So that was way easier because you can like pull up the doc on both sides." "Probably once past the learning curve that I would rather just use AI."
	Simplicity	"There's really no, unless you run into a major issue, like copy and pasting or originally writing the prompt, it's pretty, you know, pretty straightforward."
Behavioral intention to use	General preference	"I was very pleasantly. Very pleasantly and like it [AI]." "If this was about you know, volume and things like that, then yes, of course I would absolutely use AI to assist with that."
	Conditional preference	"I think it'll be helpful, but until it is shown, you know, through studies or whatever, however we want to verify the AI, until it has shown that it can do a really good job, very consistently. then I can't place that much trust on it."
	Content-dependent preference	"So things that can be automated, like, number to number, I can trust that, but if you're taking a paragraph of words and then computing that, I feel that there may be details that are being lost or something not translated."
Perceived liability of usage	Responsibility attribution	"I think as a doctor, we are responsible [for], all the work and interpretation." "Because everything is always going to be my responsibility. But if AI is advertising its ability to do a certain thing and it makes a mistake, that AI also needs to take responsibility, not take my responsibility, but take its own responsibility and correcting it and developing its AI in a way that doesn't make that mistake again." "Yeah, I have no idea who should be responsible. [...] If you extract all the data through AI, [...] because usually, like as an undergrad, it was me [dealing] with data collection. So in that case, if the data's collected wrong, you wouldn't really blame the undergrad students. It would be the doctor. Or it'd be the PI. I don't know."

Themes	Subthemes	Representative quotes
	Perceived ownership of extraction	"It might do it faster so it's more convenient. And I was like, well, maybe that indirectly helps me understand it because it's quicker. But I'm like, the understanding is still coming from me. Whether I manually do it or I have AI do it, I'm still interpreting the results. " "I don't think using AI would compromise my ability to make decisions about a case because it's not changing the information about the case, it's organizing the information. So for me, I don't think that would make me. necessarily change how I would treat something, whether I had the information presented to me regular or via AI."
	Post-AI verification behavior	"And so I think even if I was repeatedly getting my results in their results the same every time, I think it's healthy for me to always have a, like a caution towards letting it do that for me. Checking its accuracy." "I do think it's important to not just throw everything into the AI and trust it. So I still want to go through it."
AI in veterinary education	Application	"Patient interaction and clinical reasoning. Like, we don't get a lot of opportunity one-on-one to go through clinical reasoning, and this is, like a pseudo form of that. Because I feel like a lot of what we do in school is groups. And so it's really nice because you get more perspectives, but you're not testing yourself as much because you have several minds working on the same thing versus like me trying to figure it out on my all by myself." "I got to try out an AI simulation for anesthesia. And I think that could be extremely useful. Before we do like a live imitation and live anesthesia. I think that would be a phenomenal tool because in this particular experience, [...] it just kind of sets you in the moment, so I do think like AI simulation would be really helpful. I don't know for like. I don't know about like surgery or anything like that, but at least for classes where it is kind of a pressure cooker situation, but you could do it yourself or it's like a self-study type of a thing." "[...] using that VetGeni in a lot of our communication skills exams, and I think that we should be using that in all of it and whenever there's part in a curriculum with recounting physical exams and histories, I think that should be more AI involved."
	Workforce expectation	"Because I think what if you leave vet school and your practice is only using AI notetaking, like you only diction within your client and then you don't know how to use it. You're not really day one ready." "It's inevitable, so it's like, there's not much you could do about it. It's gonna take over a lot of sectors. and like I wouldn't say like replace jobs, but it's definitely going to replace a lot about how work is done."
	Benefits	"And it's going to help a lot with work life balance, I think, because they talk about, I hear a lot of veterinarians. They complain about the end of the day after the cases, they have to write out all their soaps, and that's what keeps them staying after hours. But with

Themes	Subthemes	Representative quotes
		AI, all you have to do is put that all in, double check it, and you're good to go."

Discussion

Overall, ChatGPT was 34 times faster, with accuracy over 90%, comparable to veterinary students' performance in medical data extraction. Student perceptions from both the survey and interviews were generally positive toward AI, aligned with its better performance on extraction accuracy and efficiency. The mixed-methods design, including cross-examination of extraction accuracy, time spent, and participant viewpoints, revealed the factors underlying the observed performance differences as discussed below.

When examining the time spent in each case, we noticed a sequential drop in manual extraction time from Case 1 to 3, suggesting a learning effect, as students tend to complete tasks in order, while AI remained relatively stable (Fig. 2). This difference shows how AI bypassed the human learning curve and reached peak efficiency from the first case.

Regarding extraction quality, non-compliance errors were made by both human and AI, but only found in the AI-assisted task. One student submitted identical output across all three cases, likely due to human oversight. In another student's AI output of Case 3, the AI returned phrases instead of pre-defined codes, suggesting that it understood the task but failed to follow strict output instructions. Unlike formatting issues, which affected entire cases, other error types occurred sporadically at the level of individual data points. Students made more numerical errors during manual extraction than did AI, a vulnerability that participants attributed to repetition and attention fatigue. Similarly, students made more errors than AI in categorical data, which appeared to require greater cognitive engagement. Unlike numerical values, which are often presented in relatively standardized formats, categorical findings may be dispersed across the medical record and surrounded by irrelevant background information. Prior work has similarly observed such likelihood of human extractors missing critical details during manual chart review²⁶. This challenge manifested as extraction errors in clinical signs (present or absent). One participant described this burden as follows:

If you have a whole bunch of random ones from a whole bunch of different places, then you're like, hopefully I can find the answer. [...] If you have to read the whole thing to find your two sentences that are important, I feel like that could be pretty mentally demanding.

Another aspect associated with clinical narratives was ambiguity and the potential for over-inference. This phenomenon was observed when extracting hydration status and categorizing effusion, where records did not always provide an explicit determination. Extractors reported difficulty parsing the original author's intent. Inconsistencies in medical documentation are well-recognized in specialties like

radiology²⁷ and pathology²⁸⁻³⁰. Our findings demonstrate that downstream extractors remain inherently subject to these inconsistencies in the source documentation.

Interestingly, participant perspectives varied on how extraction tasks should be divided between humans and AI. One participant suggested that AI may be overall well suited to make “black-and-white” calls for tasks such as data extraction, whereas another proposed that AI could handle numerical information while humans remain responsible for text-based extraction and interpretation. The latter preference for human judgment in narrative tasks echoed a prior study in veterinary education, where 70.4% of surveyed students agreed that human decisions should prevail over AI outputs¹⁹. However, these survey findings should be interpreted cautiously, as perceptions of AI's appropriate role are likely task-dependent and often ill-defined, unlike in the present study.

Extended from perceived ownership, accuracy, and liability, these factors appear to cluster and mutually influence one another. A natural subsequent behavior is output verification. Although participants were not explicitly instructed to verify or not to verify the results, their behavior suggested that the study's time constraints may have discouraged them from doing so during the extraction task. Nevertheless, outside of this specific task, students generally reported a habit of verifying AI outputs, consistent with the survey finding that the majority of veterinary students believed AI-generated responses should be verified¹⁷:

I was always referencing to actual trustworthy material.

Ultimately, accuracy defines data extraction quality. Benchmark “ground truth” is typically established through domain expert annotations. However, the benchmark only reflects a subjective consensus, rather than recovering a single, objective “truth” from medical records. As more human annotators contribute diverse clinical perspectives, the benchmark itself continually recalibrates, illustrating that human-derived ground truth is an inherently moving target³¹⁻³³. Recognizing this distinction matters when interpreting accuracy metrics for both human- and AI-based extraction systems. In this study, ChatGPT naturally defaulted to a conservative extraction approach similar to domain-expert standards, contrasting with common perceptions of generative AI's tendency to hallucinate³⁴⁻³⁶. This behavioral difference may be coming from the prompt's requirement for AI to quote the original record, indirectly making its extractions more literal. This suggests that adding explicit instructions to human extractors (e.g., avoid over-interpretation) could potentially close the accuracy gap.

In this classroom-based study, AI education in veterinary medicine served as the foundational context of the work. While the task focused on data extraction, students described broader applications of AI across the veterinary field, such as AI scribes, clinical simulations, and self-study tools. As highlighted in the Results, these attitudes were shaped by workforce expectations, leading students to advocate for more exposure to AI and a more open and supportive environment:

I do think it needs to be more talked about, and I don't think that it should be so polarized.

At the same time, students expressed caution regarding the timing and implementation of these tools:

And I think that's where you're going to hit a point in education where if you use it to assist students too much and they don't learn how to do it, [let's] say in the medical record, if they don't learn how to extract the data manually first, and only ever used AI, then if they end up in a situation where they need to extract the data or they need to validate the AI's side, they don't know how to do that just because they never did it.

These findings indicate that veterinary students support greater AI integration into education, but under the conditions of open discussion and cautious introduction. As AI becomes more prevalent in clinical workflow, student perspectives can help inform future curriculum development in veterinary programs.

Overall, veterinary students favored AI for its efficiency and ability to reduce manual workload, despite noting potential technical challenges in the workflow. Students recognized clinical understanding and verification as critical to output quality, while noting that the time required for these steps came at the expense of some efficiency gains. Ultimately, these perceptions shaped their behavioral intention to use AI in clinical data extraction.

This study has several limitations. First, while the sample size for qualitative interviews was adequate to achieve thematic saturation³⁷, pairing these interviews directly with survey responses resulted in a small cohort, limiting generalizability and further subgroup analysis. Second, the absence of a verification step for extracted data introduces uncertainty in the overall assessment of accuracy and efficiency, representing an important direction for future investigation. Third, the survey instrument was not formally validated through pilot testing, which may have resulted in variation in how individual questions were interpreted. For example, participants interpreted the term “tech-savviness” differently, an issue that became apparent during the subsequent semi-structured interviews. Future studies should pilot test survey instruments before administration. Fourth, data extraction is not limited to veterinary students. Undergraduate students, for example, may also be involved, but given veterinary students’ familiarity with veterinary medicine, a time difference between manual and AI may be more evident in less experienced extractors.

Conclusions

This study is the first to measure the accuracy and time required for manual and AI-assisted extraction of veterinary medical records. ChatGPT completed extraction about 34 times faster at 15 seconds per case, compared with about 8.5 minutes for veterinary student; both approaches achieved more than 90% extraction accuracy. ChatGPT excelled at reducing the repetition of manual data extraction, yet both humans and AI generated errors when navigating lengthy narratives and ambiguous documentation requiring nuanced interpretation.

By extending the TAM to include perceived liability and prior experience, this study found that student acceptance of AI is driven by both performance and by concerns about accountability and familiarity

with the technology.

In summary, ChatGPT shows strong potential as an efficient tool to support veterinary clinical data extraction, with human oversight remaining essential for information that is ambiguous or heavily dependent on clinical context.

Declarations

Competing Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was partially funded by Generative AI Research for Education (GAIRE) program of The Texas A&M Institute of Data Science.

Author Contribution

YC developed the methodology and ground truth, performed data curation and analysis, conducted the qualitative coding, and wrote the original draft. CPC contributed to overall study design, qualitative coding, and manuscript review. AS led the qualitative coding process and contributed to manuscript review. All authors reviewed and approved the final manuscript.

Acknowledgement

The authors thank all participating students for their contributions and Drs. Jonathan Lidbury and Avery Huang for establishing the ground truth of extracted data. We also thank Dr. Sheila Cherry for reviewing the manuscript and providing editorial feedback.

Data Availability

The data are available from the authors upon reasonable request.

References

1. Farrell, S., Appleton, C. & Noble, P. J. M. Al Moubayed, N. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Sci. Rep.* **13**, 18015 (2023).

2. Noble, P. J. M., Appleton, C., Radford, A. D. & Nenadic, G. Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. *PLoS One*. **16**, e0260402 (2021).
3. Conroy, M., Brodbelt, D. C., O'Neill, D., Chang, Y. M. & Elliott, J. Chronic kidney disease in cats attending primary care practice in the UK: a VetCompass™ study. <https://doi.org/10.1136/vr.105100> doi:10.1136/vr.105100.
4. Hur, B. A., Hardefeldt, L. Y., Verspoor, K. M., Baldwin, T. & Gilkerson, J. R. Describing the antimicrobial usage patterns of companion animal veterinary practices; free text analysis of more than 4.4 million consultation records. *PLoS One*. **15**, e0230049 (2020).
5. Wulcan, J. M. et al. Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. *Front Vet. Sci* **11**, (2025).
6. Reagan, K. L., Reagan, B. A. & Gilor, C. Machine learning algorithm as a diagnostic tool for hypoadrenocorticism in dogs. *Domest. Anim. Endocrinol.* **72**, 106396 (2020).
7. Kim, Y. et al. Machine learning-based risk prediction model for canine myxomatous mitral valve disease using electronic health record data. *Front. Vet. Sci.* **10**, 1189157 (2023).
8. Sobkowich, K. E. Demystifying artificial intelligence for veterinary professionals: practical applications and future potential. *Am. J. Vet. Res.* **86**, S6–S15 (2025).
9. Alzu'bi, A. A., Watzlaf, V. J. M. & Sheridan, P. Electronic Health Record (EHR) Abstraction. *Perspect. Health Inf. Manag.* **18**, 1g (2021).
10. Chang, Y. & Chu, C. P. Applying artificial intelligence to automate clinical data extraction from veterinary electronic health records: A practical guidance and comparative scoping review. *Vet. Pathol.* **03009858261461761** 10.1177/03009858261461761 (2026).
11. Lan, Z. & Turchin, A. Impact of possible errors in natural language processing-derived data on downstream epidemiologic analysis. *JAMIA Open*. **6**, ooad111 (2023).
12. Yisha, Z. et al. Data extraction error and its implications on systematic reviews in urology: a protocol. *Int. J. Surg. Protoc.* **29**, 102–107 (2025).
13. Hur, B., Wang, L. L., Hardefeldt, L. & Yetisgen, M. Is that the right dose? Investigating generative language model performance on veterinary prescription text analysis. in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing* (eds Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K. & Tsujii, J.) 390–397 Association for Computational Linguistics, Bangkok, Thailand, (2024). 10.18653/v1/2024.bionlp-1.30
14. Pinard, C. J. et al. Precision in Parsing: Evaluation of an Open-Source Named Entity Recognizer (NER) in Veterinary Oncology. *Vet. Comp. Oncol.* **23**, 102–108 (2025).
15. Ge, J., Li, M., Delk, M. B. & Lai, J. C. A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record. *Gastroenterology* **166**, 707–709e3 (2024).
16. Chen, Y. et al. Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system. *BMC Med. Res. Methodol.* **22**, 136 (2022).

17. Hong, J. C., Fairchild, A. T., Tanksley, J. P., Palta, M. & Tenenbaum, J. D. Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts. *Jamia Open*. **3**, 513–517 (2020).
18. Oropesa, A. L., Mendo-Lázaro, S., Gonzalez, A., León-Del-Barco, B. & Tapia, J. A. A. Artificial intelligence in veterinary education: self-perceived knowledge, use, and attitudes among veterinary students in Spain and Portugal. *Front Vet. Sci* **13**, (2026).
19. de Brito, C. et al. Integrating artificial intelligence into veterinary education: student perspectives. *Front Vet. Sci* **12**, (2025).
20. Reagan, K. L., Boudreaux, K. & Keller, S. M. Veterinary students exhibit low artificial intelligence literacy but agree it will be deployed to improve veterinary medicine. (2025). <https://doi.org/10.2460/ajvr.25.03.0082> doi:10.2460/ajvr.25.03.0082.
21. Rhind, S., Connelly, L., Baillie, S., Shaw, D. J. & Dickson, J. Use of Generative AI by Veterinary Students in Two UK Vet Schools: A Moment in Time. *J Vet. Med. Educ* **0**, e20250078 .
22. Davis, F. D. & Perceived Usefulness Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **13**, 319–340 (1989).
23. Holden, R. J. & Karsh, B. The technology acceptance model: its past and its future in health care. *J. Biomed. Inf.* **43**, 159 (2010).
24. Huang, Y. T. & Chu, C. P. P. Curriculum framework for artificial intelligence literacy in veterinary education. *Front Vet. Sci* **13**, (2026).
25. Cervera, M., Albert, M., Torres, V. & Pelechano, V. On the usefulness and ease of use of a model-driven Method Engineering approach. *Inform. Syst.* **50**, 36–50 (2015).
26. Yin, A. L. et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *Int. J. Med. Inf.* **157**, 104622 (2022).
27. Weissman, A. et al. A Survey of Radiologists and Referring Veterinarians Regarding Imaging Reports. *Vet. Radiol. Ultrasound*. **57**, 124–129 (2016).
28. Giglia, G. et al. Modifying phrases in veterinary pathology diagnostic reports: The veterinary professionals' perception on the diagnostic confidence. *Vet. Pathol.* **62**, 963–971 (2025).
29. Christopher, M. M. & Hotz, C. S. Cytologic diagnosis: expression of probability by clinical pathologists. *Vet. Clin. Pathol.* **33**, 84–95 (2004).
30. Rishniw, M. & Freeman, K. P. Clinical pathologists should limit modifier terms used to denote probability of a diagnosis: a survey-based study. (2023). <https://doi.org/10.2460/javma.22.11.0488> doi:10.2460/javma.22.11.0488.
31. Syololypavan, A., Sleeman, D., Wu, H. & Sim, M. The impact of inconsistent human annotations on AI driven clinical decision making. *npj Digit. Med.* **6**, 26 (2023).
32. Nelson, S. J. et al. Are ICD codes reliable for observational studies? Assessing coding consistency for data quality. *Digit. Health*. **10**, 20552076241297056 (2024).

33. Luitse, D. & Blanke, T. Zones of (un)certainity: studying the temporal stabilization of ground truths for medical AI. *AI Soc.* <https://doi.org/10.1007/s00146-026-03228-x> (2026).

34. Roustan, D. & Bastardot, F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interact. J. Med. Res.* **14**, e59823 (2025).

35. Gondode, P., Duggal, S. & Mahor, V. Artificial intelligence hallucinations in anaesthesia: Causes, consequences and countermeasures. *Indian J. Anaesth.* **68**, 658–661 (2024).

36. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **8**, 274 (2025).

37. Littlewood, K. & Gardner, D. A brief guide to qualitative research in veterinary science: interviews, focus groups, surveys and reflexive thematic analysis for practitioners and researchers. *N Z. Vet. J.* **74**, 339–349 (2026).

Figures

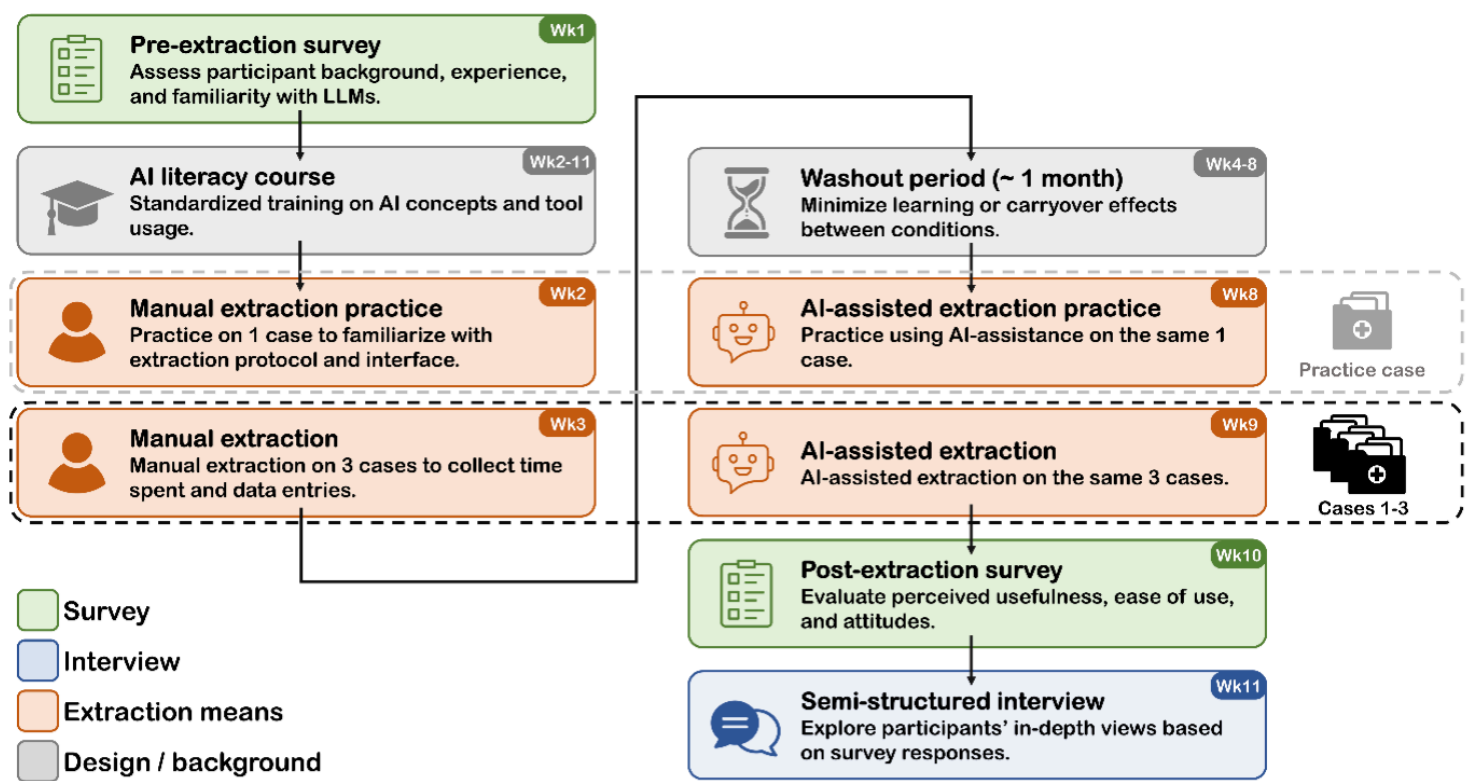


Figure 1

Study workflow. LLM, large language model. AI, artificial intelligence. Wk, Week.

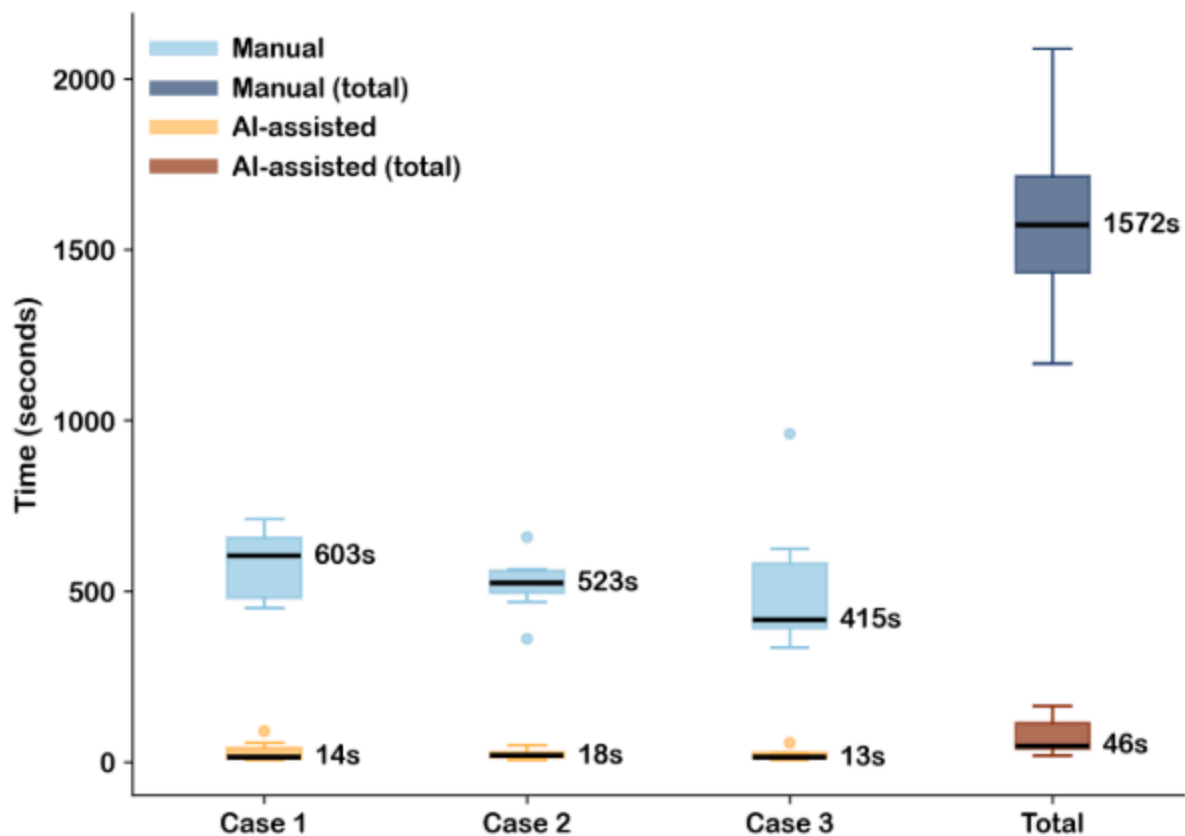


Figure 2

Extraction time differences between manual and AI-assisted extraction. Extraction time (seconds) for manual (blue) and AI-assisted (orange) tasks across three cases (Cases 1–3) and aggregated total (darker shades). The horizontal line within each box indicates the median, with its value annotated adjacent to each box. Boxes show the interquartile range (IQR). Whiskers extend to 1.5 times the IQR, with individual points representing outliers.

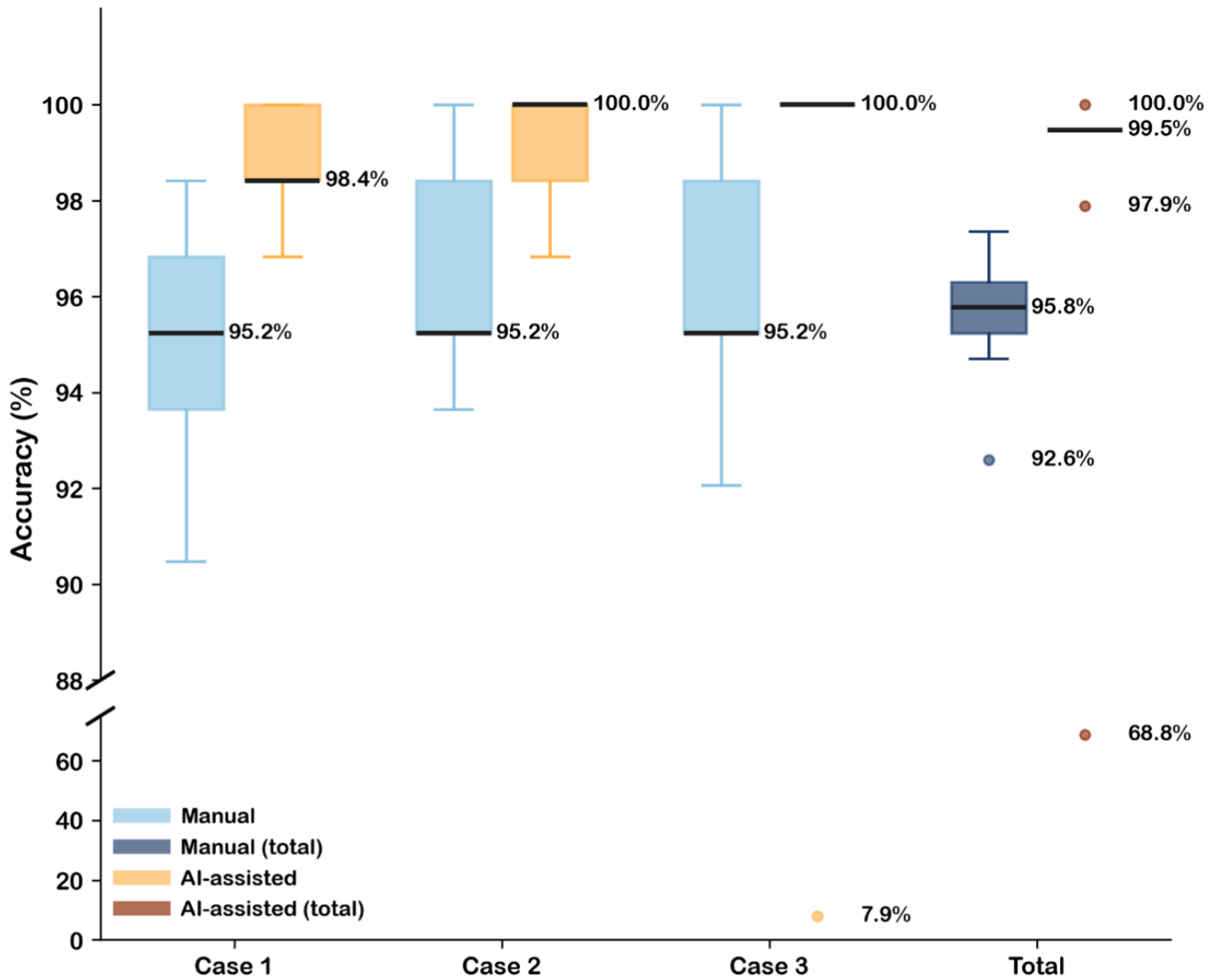


Figure 3

Extraction accuracy comparison between manual and AI-assisted extraction. Extraction accuracy for manual (blue) and AI-assisted (yellow) tasks across three cases (Cases 1–3) and aggregated total (darker shades). The horizontal line within each box indicates the median, with its value annotated adjacent to each box. Boxes show the interquartile range (IQR). Whiskers extend to 1.5 times the IQR, with individual points representing outliers.

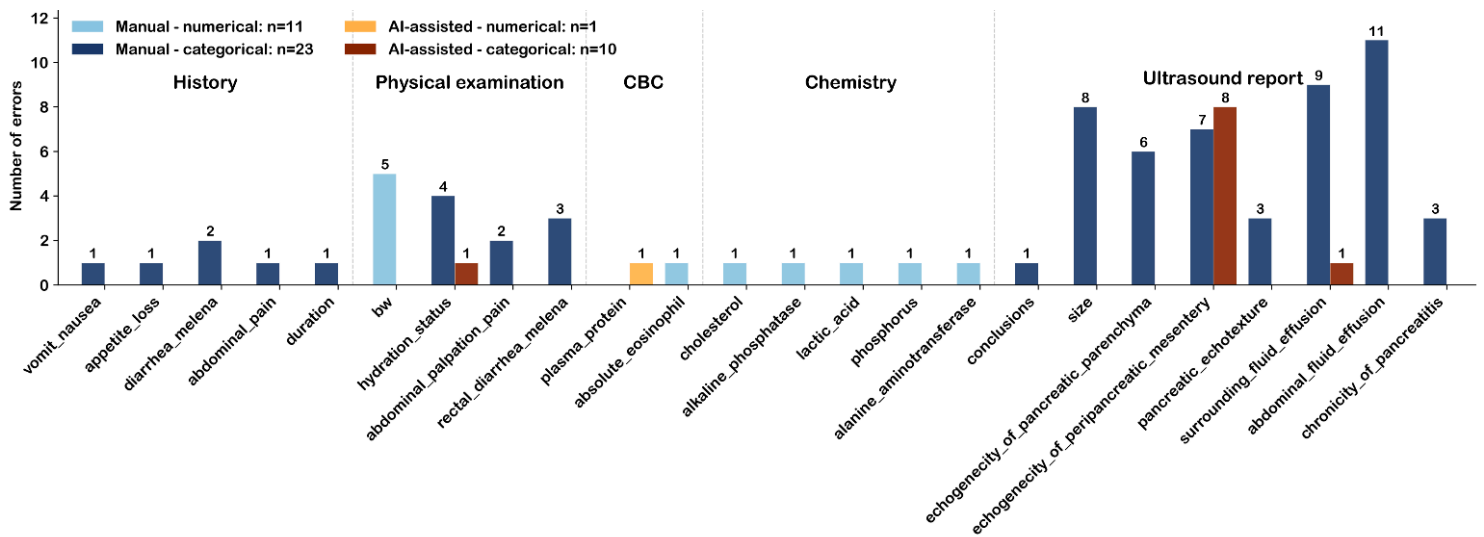


Figure 4

Distribution of extraction error. Error counts per variable for both manual (blue) and AI-assisted (orange) extractions. Darker shades represent categorical data type; lighter shades represent numerical data type. Bars are grouped by clinical record sections. The x-axis label is centered between two bars (AI-assisted on the right, manual on the left). A missing bar indicates no errors. Values of n in the legend represent the total sum of errors. CBC, complete blood count. bw, body weight.

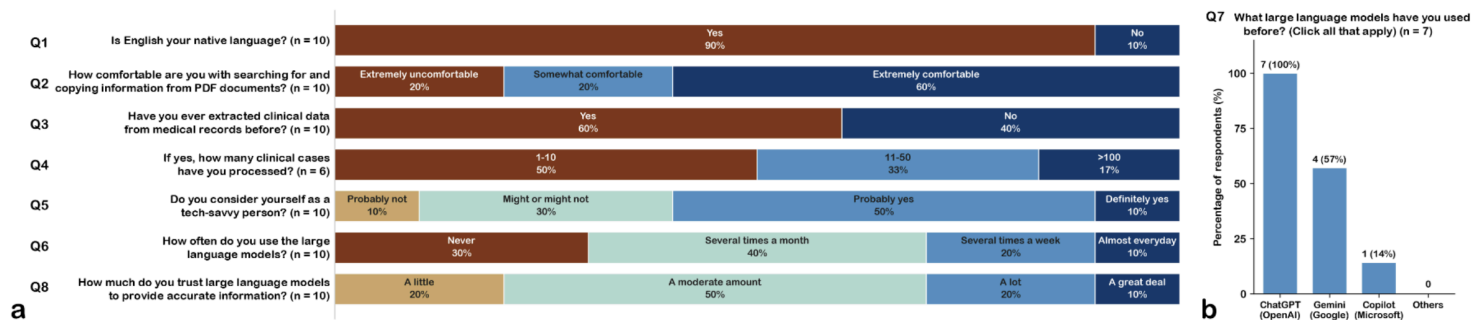


Figure 5

Pre-extraction survey responses. Distribution of responses across eight survey items assessing prior extraction experience, technological proficiency, and large language models (LLMs) familiarity and trust. (n = 10) **(a)** Q1 and Q3 are binary. Q4 is an ordinal follow-up for experienced participants from Q3 (n = 6). Q2, Q5, Q6, and Q8 used 5-point Likert scales. Response categories with 0% are not shown. **(b)** Self-reported prior use of specific LLMs among respondents who indicated usage from Q6 (n = 7). As multiple responses were allowed, percentages exceeded 100%.

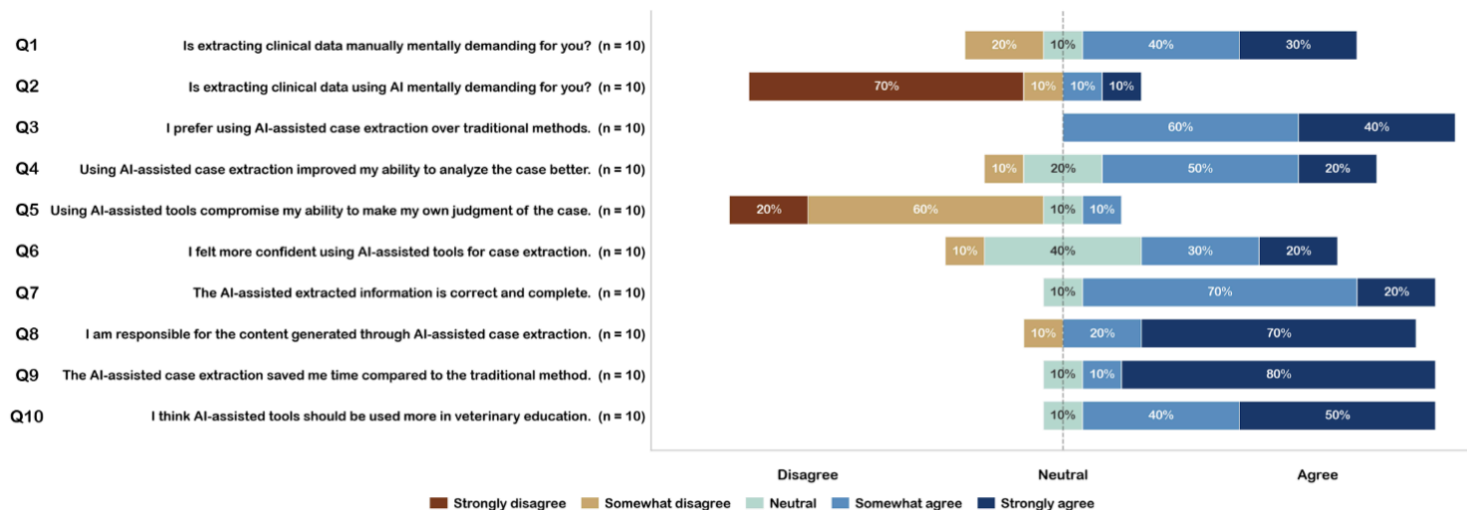


Figure 6

Post-extraction survey. Distribution across 10 items assessing participants' (n = 10) attitudes toward comparative extraction. The survey measured perceived usefulness (Q4, Q5, Q7, Q9), perceived ease of use (Q1, Q2), behavioral intention to use (Q3, Q6), perceived liability (Q8), and attitudes toward AI in veterinary education (Q10). All items used 5-point Likert scales. Response categories with 0% are not shown. AI, artificial intelligence.

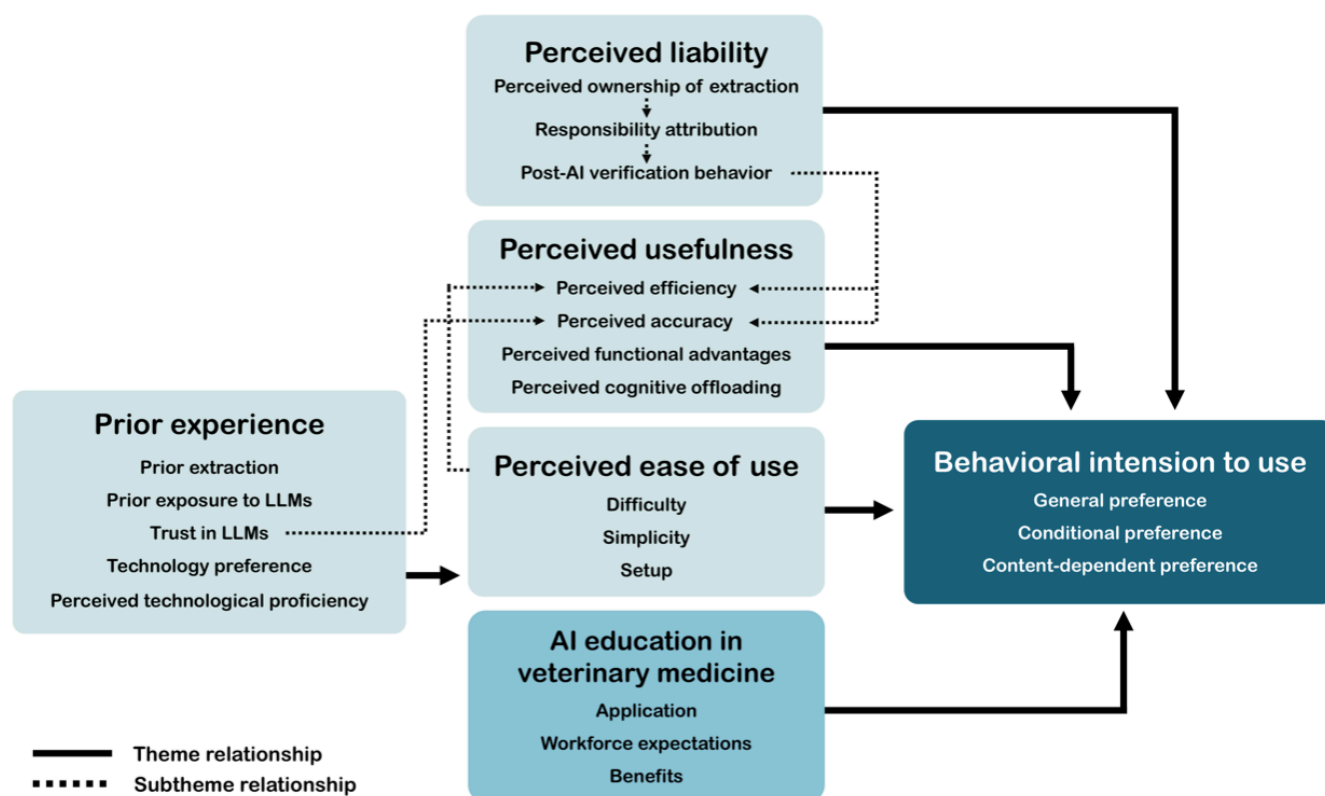


Figure 7

Thematic framework of student perceptions toward AI-assisted clinical data extraction. Themes and subthemes are organized into each construct. Within each box, themes are displayed in larger bold text and subthemes in smaller text. AI, artificial intelligence, LLMs, large language models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary20260818.pdf](#)