scientific reports

OPEN

Preliminary evaluation of ChatGPT model iterations in emergency department diagnostics

Jinge Wang[1], Kenneth Shue[1], Li Liu[2,3] & Gangqing Hu[1]✉

Large language model chatbots such as ChatGPT have shown the potential in assisting health professionals in emergency departments (EDs). However, the diagnostic accuracy of newer ChatGPT models remains unclear. This retrospective study evaluated the diagnostic performance of various ChatGPT models—including GPT-3.5, GPT-4, GPT-4o, and o1 series—in predicting diagnoses for ED patients ($n = 30$) and examined the impact of explicitly invoking reasoning (thoughts). Earlier models, such as GPT-3.5, demonstrated high accuracy for top-three differential diagnoses (80.0% in accuracy) but underperformed in identifying leading diagnoses (47.8%) compared to newer models such as chatgpt-4o-latest (60%, $p < 0.01$) and o1-preview (60%, $p < 0.01$). Asking for thoughts to be provided significantly enhanced the performance on predicting leading diagnosis for 4o models such as 4o-2024-0513 (from 45.6 to 56.7%; $p = 0.03$) and 4o-mini-2024-07-18 (from 54.4 to 60.0%; $p = 0.04$) but had minimal impact on o1-mini and o1-preview. In challenging cases, such as pneumonia without fever, all models generally failed to predict the correct diagnosis, indicating atypical presentations as a major limitation for ED application of current ChatGPT models.

Keywords ChatGPT, Large Language models, Emergency medicine, Diagnosis, Model iterations

Application of AI in medicine has shown great promise in advancing healthcare outcomes. As of Feb 2025, the United States Food and Drug Administration has approved 1,016 artificial intelligence and machine learning-enabled medical devices spanning 17 specialties, including cardiovascular, gastroenterology and radiology[1]. Moreover, a recent review of randomized controlled trials on AI interventions found that 81% of the trials reported positive improvements in their primary endpoints[2]. Along with these encouraging advancements, the recent emergence of generative AI models, including large language models (LLMs), has created new opportunities to enhance patient care particularly in high-pressure environments such as emergency departments (EDs). On one hand, generative AI models enhanced with domain-specific knowledge excel in various medical tasks, as seen with PathChat for pathology analysis[3], Flamingo-CXR for chest radiographs[4], and Med-Gemini for general medical inquiries[5]. On the other hand, LLMs trained without discipline-specific restrictions also demonstrate strong performance in handling specialized tasks, including those commonly encountered in emergency medicine. For instance, ChatGPT achieves diagnostic performance comparable to healthcare professionals[6,7]. This chatbot, along with its competitor Gemini (formerly known as Bard), also demonstrates modest accuracy in triaging patients based on severity[8–14]. Similarly, case studies highlight ChatGPT's ability to guide patients in determining whether to seek urgent care based on symptoms[15]. However, ChatGPT often errs on the side of caution, over-recommending emergent care visits[16] and over-prescribing interventions[17]. All the studies demonstrate the promise of ChatGPT in facilitating patient care in EDs yet also underscore limitations and challenges for future model refinements.

The field of LLM-based chatbots is evolving rapidly[18]. Since its launch in December 2022, ChatGPT has undergone multiple significant updates, including the GPT-3.5, GPT-4, GPT-4o, and o1 models. Much of the existing research on chatbots in EDs is based on earlier iterations, such as GPT-3.5 and GPT-4. Beyond architectural improvements, updates to training datasets also influence model performance. These developments raise critical questions about how successive model iterations influence diagnostic accuracy, a topic not yet explored in ED visits. This retrospective study addressed these gaps by evaluating the diagnostic performance of various ChatGPT models on ED diagnostic tasks, using a previously published dataset derived from real-word, de-identified patient data[6].

[1]Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, WV 26506, USA. [2]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA. [3]Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA. ✉email: michael.hu@hsc.wvu.edu

Materials and methods

Benchmark dataset

The dataset from Ten Berg et al.[6], comprising 30 patient cases, was utilized to evaluate the diagnostic performance of different ChatGPT models in ED settings. Each patient case included a single confirmed discharge diagnosis, which served as the ground truth for evaluation. The input to the ChatGPT models consisted of physician note, which captures information typically available at the time of initial ED presentation. These notes included details such as signs and symptoms, medical history, and physical examination findings. To reflect data typically available at the time of initial presentation, laboratory results were excluded. The cases were indexed as case 1 to case 30 the same as in the previous publication[6].

GPT models and testing protocol

We included all GPT models where Application Programming Interface (API) access was available at the time of experiments (September 2024). The GPT-3.5 series included gpt_3.5_turbo_1106 and gpt_3.5_turbo_0125, though their web versions are no longer accessible. The GPT-4 series included gpt_4_0314, gpt_4_0613, gpt_4_1106_preview, gpt_4_0125_preview, and gpt_4_turbo_2024_04_09. The 4o series included 4o_mini_2024_07_18, 4o_2024_05_13, 4o_2024_08_06, and chatgpt_4o_latest. The o1 series included o1-mini and o1-preview. A summary of all the models is shown in Table 1. All models were accessed through the API, with the temperature parameter set at 0.7 to mimic web-based usage condition, which balances reproducibility (lower temperature) and exploratory reasoning (higher temperature).

Two testing modes were implemented. In mode 1, GPT models were prompted to generate and rank the top three differential diagnoses without disclosing reasoning process. In mode 2, the GPT models were prompted to provide both differential diagnoses and thoughts for the reasoning. The performance of the models in these two modes was compared to assess whether asking for thoughts improved diagnostic accuracy or not. Each patient case was presented to each model three times via API. Diagnostic performance was measured based on two metrics: (1) the % of inclusion of the correct diagnosis within the top-three differential diagnoses (cases with correct diagnosis in top three / total cases) and (2) the % of identification of the correct diagnosis as the leading diagnosis (cases with correct diagnosis ranked first / total cases). Detailed prompts used for these tests are provided in **Supplementary Table S1**.

To determine accuracy, we used keyword matching while also considering synonyms. For example, if the correct diagnosis is renal colic, responses such as ureterolithiasis or renal stone were also considered correct. Due to the binary nature of keyword matching, multiple raters were not required. This approach aligns with the previous work[6], where we sourced the dataset.

Ablation test

During the evaluation, GPT models generally failed to include pneumonia as a diagnosis for three specific cases (case 1, case 26, and case 28; **Supplementary Table S2**). Further inspection revealed that the corresponding physician notes stated no fever and no chills, whereas pneumonia typically presents with a high fever. To determine whether the lack of high fever influences GPT's pneumonia diagnosis and to assess whether modifying these inputs could improve diagnostic outcomes, an ablation test was conducted. Specifically, we removed phrases such as "no fever" and "no chills" from the three physician notes and instead included a body temperature of 39 °C. The modified physician notes were re-evaluated using chatgpt_4o_latest, following the same prompt structure as the original cases asking for top three diagnosis.

Statistics

Pairwise comparisons of GPT models were conducted using *pairwise.t.test* from the *R Stats Package* with Bonferroni corrections, which controlled for Type I errors in multiple testing by adjusting the raw p-values based on the number of comparisons. For accuracy comparisons between the two testing modes—mode 2

Model Family	Model Description	Model Name	Training Data Cut-off
GPT-3.5	The GPT-3.5 Turbo family offers fast and accurate models optimized for chat and text generation, with support for large context windows and various tasks.	gpt-3.5-turbo-1106 gpt-3.5-turbo-0125	Sep 2021 Sep 2021
GPT-4	The GPT-4 family consists of advanced multimodal models capable of processing text and image inputs with text outputs.	gpt-4–0314 gpt-4–0613 gpt-4–1106-preview gpt-4–0125-preview gpt-4-turbo-2024-04-09	Sep 2021 Sep 2021 Apr 2023 Dec 2023 Dec 2023
GPT-4o	The GPT-4o family includes OpenAI's most advanced and efficient multimodal models, supporting text and image inputs with text outputs.	4o-mini-2024-07-18 4o-2024-05-13 4o-2024-08-06 chatgpt-4o-latest	Oct 2023 Oct 2023 Oct 2023 Oct 2023
o1	The o1 family features advanced reasoning models trained with reinforcement learning to solve complex problems through internal chains of thought.	o1-mini-2024-09-12 o1-preview-2024-09-12	Oct 2023 Oct 2023
Model information is referred from: https://platform.openai.com/docs/models			

Table 1. Overview of ChatGPT models evaluated in the study.

(with reasoning) vs. mode 1 (without reasoning)—within each GPT model, a single comparison was conducted per model, making adjustment for multiple comparisons unnecessary. Thus, a plain student t-test was used to calculate p-values.

Results

Case overview

Our study evaluated GPT models on 30 ED cases sourced from a previous study[6], which did not provide demographics information. Respiratory and gastrointestinal conditions compromised 36.7% and 23.3% of the patient cohort, respectively, with pneumonia, exacerbation of COPD/asthma, and pancreatitis being the most frequently observed diseases (Table 2). A total of 13 cases were identified where healthcare professionals from the original study misidentified the leading diagnosis[6]. These cases were categorized as "difficulty cases" here (Table 2).

Performance on top three diagnoses

We first evaluated the accuracy of each model in including the correct diagnosis within the top three differential diagnoses. Notably, the early GPT-3.5 model with an updated training dataset, gpt_3.5_turbo_1106, achieved the highest accuracy at 80.0%. In contrast, more recent models demonstrated slightly lower accuracies: chatgpt_4o_latest (78.9%) and o1_preview (74.4%) (Fig. 1a).

Next, we examined whether requesting explanations alongside diagnoses would enhance accuracy. For models developed before the 4o series, the addition of asking for thoughts did not improve diagnostic performance (Fig. 1b). However, newer 4o models showed improvements with reasoning. For instance, the accuracy of 4o_2024_05_13 increased from 76.7 to 83.3%, achieving the highest accuracy among all models. Similarly, o1_preview improved from 74.4 to 78.9%.

Performance on leading diagnosis

We also assessed the accuracy of each model in identifying the correct leading diagnosis. Models released after July 2024 generally outperformed earlier versions in this metric. The highest accuracies were achieved by chatgpt_4o_latest (60.0%) and o1_preview (60.0%), showing substantial improvements over gpt3.5-turbo-0125 (47.8%) (adjusted $p < 0.01$ for both comparisons) (Fig. 2a). Consistent with the findings for top-three differential diagnoses, requesting thoughts generally improved the accuracy of the 4o models. Specifically, 4o_2024_05_13 increased from 45.6 to 56.7% ($p = 0.03$; t-test), 4o_mini_2024_07_18 improved from 54.4 to 60.0% ($p = 0.04$), 4o_2024_08_06 increased from 56.7 to 62.2% ($p = 0.14$), and chatgpt_4o_latest improved from 60.0 to 67.8% ($p = 0.07$) (Fig. 2b). Interestingly, the o1 models (o1-mini and o1-preview) showed no improvement when asking for thoughts. As an example, **Supplementary Table S3** presents diagnostic reasoning from chatgpt_4o_latest and o1-preview for a patient with diverticulitis (case 3 from Ten Berg et al.[6]).

Performance on cases stratified by difficulty

The original study by Ten Berg et al.[6] includes 13 cases in which health professionals misidentified the leading diagnoses, which we categorized as "difficult cases" (Table 2). We evaluated the performance of the GPT models specifically for these challenging cases. Overall, model performance remained limited, with the best model—chatgpt_4o_latest—achieving an accuracy of only 38.5% for leading diagnosis (Fig. 3a). Later versions of the 4o models tended to outperform GPT-3.5 and early GPT-4 models in identifying the leading diagnosis for difficult cases. However, this trend did not extend to the performance on the top three differential diagnoses (Fig. 3b).

To better understand factors contributing to the models' failures on "difficult cases", we focused on three pneumonia cases (cases 1, 26, and 28) where physician notes explicitly stated no fever. Nearly all models consistently excluded pneumonia from their differential diagnoses. To investigate the impact of body temperature on diagnostic outcomes, we modified the physician notes to indicate a body temperature of 39 °C. Under these revised conditions, chatgpt_4o_latest successfully included pneumonia in its differential diagnoses for all three cases (**Supplementary Table S2**). This finding suggests that the absence of fever was a critical factor contributing to the models' failure to diagnose atypical pneumonia accurately.

Category	Diagnoses	Count(%)
Respiratory	Pneumonia (1*, 18, 26*, 28*, 30), Exacerbation COPD/asthma (6, 14, 19, 21, 23*, 24)	11 (36.7%)
Gastrointestinal	Diverticular bleeding (2), diverticulitis (3*), pancreatitis (4*, 12, 25), gastroenteritis (15*), Biliary colic (10*)	7 (23.3%)
Infectious	Urosepsis/UTI (5, 22*), cholangitis (9*)	3 (10.0%)
Cardiovascular	Acute/chronic heart failure (8, 16), AAA (11*)	3 (10.0%)
Urogenital	Renal colic (2)	2 (6.7%)
Critical/Surgical	Ovarian torsion (13), inguinal hernia (7*)	2 (6.7%)
Other	Decompensated liver cirrhosis (17), vasculitis (20*)	2 (6.7%)
In paratheses are case ID cited from the original study[6]. *: difficult case.		

Table 2. Summary of cases.

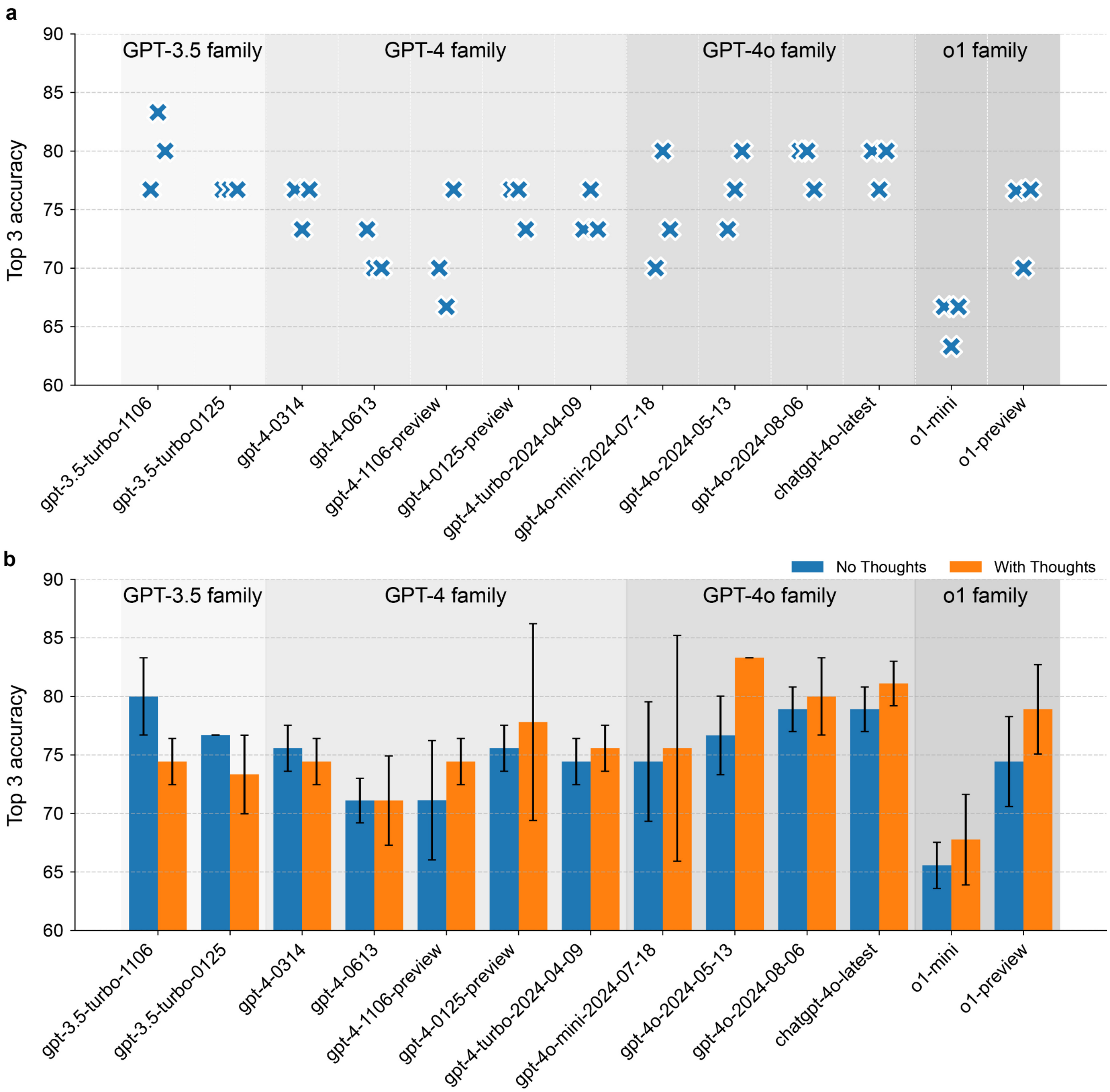

Fig. 1. Accuracy based on the top three differential diagnoses across ChatGPT models for ED cases. **(a)** Accuracy of various ChatGPT models in including the correct diagnosis within the top three differential diagnoses, evaluated without providing thoughts. **(b)** Comparison of average accuracy across models when generating top-three differential diagnoses without (blue bars) versus with asking for thoughts (orange bars). Error bars represent the standard deviation (*n* = 3).

Power analysis

Since this study utilized public data, which means that the sample size is pre-determined, a post-hoc power analysis was conducted to estimate its ability to detect meaningful differences in diagnostic accuracy. Given 30 cases per group and independent triplicate experiments, the analysis indicated ~ 80% power to detect a 20% absolute difference in accuracy at α = 0.05, assuming a baseline accuracy of 45%, which corresponded to GPT-3.5 models and early GPT4 models. These results suggest that the current sample size is sufficient to support a 20% accuracy difference observed in model comparisons—for instance, chatgpt-4o-latest vs. gpt-3.5-turbo-0125 in predicting leading diagnosis when prompted for reasoning (Fig. 2b). However, a larger sample size will be required to consolidate smaller differences between models.

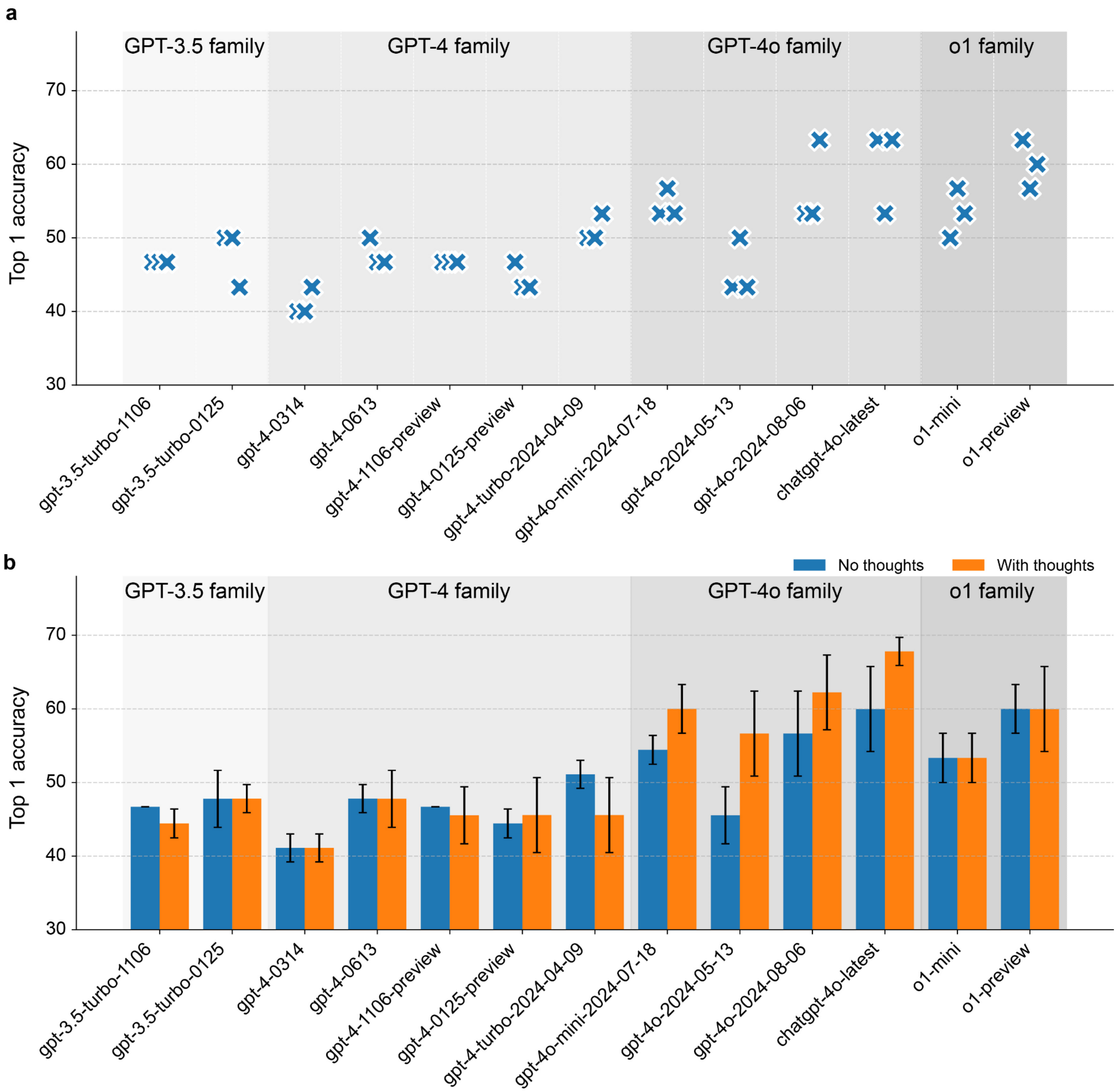

Fig. 2. Accuracy based on the leading diagnosis across ChatGPT models for ED patients. **(a)** Accuracy of various ChatGPT models in predicting the correct leading diagnosis, evaluated without providing thoughts. **(b)** Comparison of average accuracy in identifying the leading diagnosis across models when generating top-three differential diagnoses without (blue bars) versus with asking for thoughts (orange bars). Error bars represent the standard deviation ($n = 3$).

Discussion

Since its introduction in December 2022, numerous studies have evaluated ChatGPT's performance across various medical domains[19–21]. As the model continue to evolve, research has expanded to assess its performance across model iterations in areas such as medical examinations[22,23], patient record documentation[24], disease diagnosis[25,26], ECG analysis[27], and medical image interpretation[28]. Building on the work of Ten Berg et al.[6], our study focused specifically on diagnostic tasks in emergency departments by comparing the performance of various GPT model iterations. Additionally, we explored the extent to which explicitly asking for thoughts could enhance diagnostic accuracy.

Our findings indicate that accuracy in including the correct diagnosis among the top-three differential diagnoses did not consistently improve with newer model iterations. However, accuracy in prioritizing the correct diagnosis as the leading diagnosis showed notable improvement. This suggests that while newer models, particularly GPT-4o and o1, better suited for tasks requiring focused prioritization of diagnoses, older models like GPT-3.5 still hold value in scenarios that demand broader diagnostic coverage.

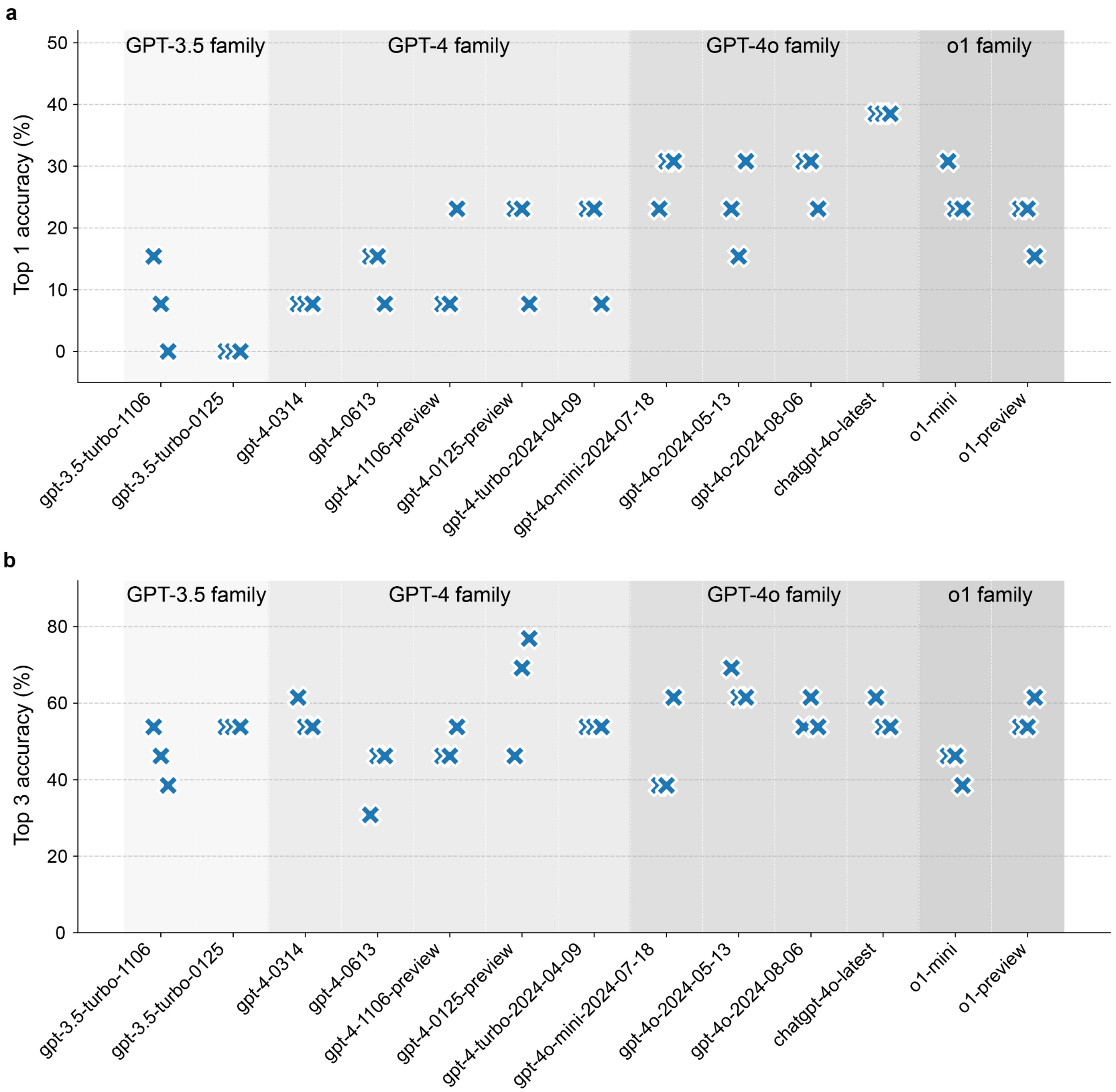

Fig. 3. Diagnosis accuracy across ChatGPT models for "difficult cases". **(a)** Accuracy of various ChatGPT models in predicting the correct leading diagnosis for difficult cases, evaluated with asking for explanations. (**b**) Accuracy of various ChatGPT models in including the correct diagnosis within the top three differential diagnoses for difficult cases, evaluated with asking for explanations. Experiments were conducted in triplicate.

Explicitly requesting output of thoughts leading to the diagnoses significantly improved the performance of the GPT-4o models but did not enhance the accuracy of the o1 models. This discrepancy may stem from differences in model architecture and training focus. Since a key update in the o1 models compared to the GPT-4o models is the integration of advanced reasoning capabilities in the structure, requesting reasoning in the prompts may not offer additional benefits for the o1 models. Future research on ChatGPT in EDs should investigate whether requesting explanations can enhance performance in other applications, such as triage and decision-making support.

While we observed a general improvement in the 4o models over earlier versions when predicting the leading diagnosis with reasoning, the lack of improvement in the o1 models was unexpected. This may stem from their architectural incompatibilities with the study's structured diagnostic tasks. Specifically, o1's native chain-of-thought reasoning is optimized for open-ended problem-solving, which may conflict with the structured, diagnosis-constrained prompts used in this study. Additionally, previous research[29,30] has shown that rigid prompting strategies can degrade the performance of self-guided reasoning models like o1, potentially by disrupting their intrinsic decision-making workflows.

The results from the ablation test on body temperature for the three challenging pneumonia cases highlighted the significant role of fever absence in the models' diagnostic inaccuracies. However, alternative explanations should also be considered. For instance, certain populations—such as older adults, infants, and immunocompromised individuals—may not exhibit fever as a symptom of pneumonia[31]. Additionally, prior use of medications such as nonsteroidal anti-inflammatory drugs (also known as painkillers or NSAIDS) before ED visits could suppress fever. These findings emphasize the importance of accounting for atypical ED presentations and external factors, which may not be adequately represented in the training data. Enhancing model performance in addressing such ED tasks may benefit from approaches such as retrieval-augmented generation[32]. Another potential direction is the integration of additional data modalities, such as laboratory results and imaging findings, to improve the model's diagnostic performance. Nevertheless, the current low performance on complex and atypical cases highlights the necessity of human oversight to ensure high-quality, patient-centered care when leveraging LLMs as assistive tools.

As future model iterations are expected to bring further performance improvements, ethical considerations for integrating GPT into clinical settings become pivotal. These include but are not limited to training data bias, data privacy concerns, and security vulnerabilities[33–35]. To address these challenges, integration should align with evolving regulatory frameworks, such as the European Artificial Intelligence Act, which mandates rigorous risk assessments, transparency, and accountability—particularly for high-stakes applications like emergency diagnostics[36]. Additionally, physicians' satisfaction on their interactions with ChatGPT—including both its diagnostic capabilities and explanatory reasoning—plays a crucial role in its potential adoption in ED settings. Conducting a survey on physician satisfaction could help identify areas for improvement in this direction, as having been recently indicated from other specialties[37].

This study has several limitations. Firstly, the small sample size ($n = 30$) may restrict the generalizability of the findings. Incorporating a broader and more diverse dataset that captures a wide range of disease presentations should be considered in future studies to validate the generalizability of the current findings. Secondly, restricting the analysis to the top three diagnoses may have narrowed the models' ability to demonstrate diagnostic breadth. Thirdly, all cases in the dataset show a single diagnosis at discharge, whereas many patients may present with multiple concurrent diagnoses. Fourthly, the ground truth relies on discharge diagnosis, which may not always be accurate[6]. Lastly, model overfitting is a valid concern. However, this is unlikely for the GPT models evaluated in our study. The cases analyzed were sourced from a subscription-based journal that is not freely accessible[6]. Furthermore, the consistently poor performance of GPT models on "difficult cases," regardless of version, suggests that this data is not likely included in the training of recent GPT models.

In summary, our evaluation of various ChatGPT models for ED diagnostic tasks yielded the following key findings: (1) Earlier models, such as GPT-3.5, exhibited high accuracy in identifying the top three differential diagnoses but underperformed in recognizing the leading diagnosis compared to newer models like ChatGPT-4o-latest and o1-preview; (2) the latest GPT-4o models—but not the o1 models—demonstrated improved performance in identifying the leading diagnosis when explicitly prompted to provide reasoning; and (3) all ChatGPT models showed limitations in diagnosing patients with atypical disease presentations. Further improvements could be achieved by integrating additional patient data modalities, such as imaging results and laboratory test findings. Additionally, new strategies, such as the development of domain-adapted GPT-based models[32,38] should be considered to address more complex clinical scenarios, including cases involving atypical presentations and multiple concurrent conditions. Moreover, randomized controlled trials will be necessary to assess the extent to which GPT can enhance physicians' diagnostic accuracy and efficiency.

Data availability

Python code for API access to different ChatGPT models are accessible at GitHub: https://github.com/JingeW/GPT-Emergency-Department-Diagnosis-Evaluation. Responses to all the 30 cases from all models under Mode 1 (diagnoses only) and Mode 2 (diagnoses with thoughts) are accessible at Mendeley: https://data.mendeley.com/datasets/vf64zj89x9/1.

Received: 3 December 2024; Accepted: 19 March 2025
Published online: 26 March 2025

References

1. United States Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. [cited 2025 Feb 12]; (2024). Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
2. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health.* **6** (5), e367–e373 (2024).
3. Lu, M. Y. et al. A multimodal generative AI copilot for human pathology. *Nature* **634** (8033), 466–473 (2024).
4. Tanno, R. et al. *Collaboration between clinicians and vision-language models in radiology report generation. Nat. Med.*, (2024).
5. Saab, K. et al. *Capabilities of Gemini Models in Medicine.* arXiv:2404.18416 (2024). https://doi.org/10.48550/arXiv.2404.18416
6. Ten Berg, H. et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Ann. Emerg. Med.* **83** (1), 83–86 (2024).
7. Hoppe, J. M. et al. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. *J. Med. Internet Res.* **26**, e56110 (2024).
8. Colakca, C. et al. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. *Sci. Rep.* **14** (1), 22106 (2024).
9. Gan, R. K. et al. Performance of Google bard and ChatGPT in mass casualty incidents triage. *Am. J. Emerg. Med.* **75**, 72–78 (2024).
10. Kaboudi, N. et al. *Diagnostic Performance of ChatGPT to Perform Emergency Department Triage: A Systematic Review and Meta-analysis.* medRxiv, : p. 2024.05.20.24307543. (2024).

11. Masanneck, L. et al. Triage performance across large Language models, ChatGPT, and untrained Doctors in emergency medicine: comparative study. *J. Med. Internet Res.* **26**, e53297 (2024).
12. Meral, G. et al. Comparative analysis of ChatGPT, gemini and emergency medicine specialist in ESI triage assessment. *Am. J. Emerg. Med.* **81**, 146–150 (2024).
13. Paslia, S. et al. Assessing the precision of artificial intelligence in emergency department triage decisions: insights from a study with ChatGPT. *Am. J. Emerg. Med.* **78**, 170–175 (2024).
14. Sarbay, I., Berikol, G. B. & Ozturan, I. U. Performance of emergency triage prediction of an open access natural Language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. *Turk. J. Emerg. Med.* **23** (3), 156–161 (2023).
15. Halaseh, F. F. et al. *ChatGPT's Role in Improving Education Among Patients Seeking Emergency Medical Treatment. Western J. Emerg. Med.*, **25**(5). (2024).
16. Salazar, G. Z. et al. *Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus J. Med. Sci.*, **15**(9). (2023).
17. Williams, C. Y. K., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large Language models to provide clinical recommendations in the emergency department. *Nat. Commun.* **15** (1), 8236 (2024).
18. Dam, S. K., Hong, C. S., Qiao, Y. & Zhang, C. *A Complete Survey on LLM-based AI Chatbots*. arXiv:2406.16937 (2024). https://doi.org/10.48550/arXiv.2406.16937
19. Fatima, A. et al. ChatGPT in medicine: A cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. *Med. (Baltim).* **103** (32), e39250 (2024).
20. Wei, Q. et al. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J. Biomed. Inf.* **151**, 104620 (2024).
21. Liu, M. et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and Meta-Analysis. *J. Med. Internet Res.* **26**, e60807 (2024).
22. Rojas, M. et al. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean medical licensing examination: observational study. *JMIR Med. Educ.* **10**, e55048 (2024).
23. Liu, M. et al. Evaluating the effectiveness of advanced large Language models in medical knowledge: A comparative study using Japanese National medical examination. *Int. J. Med. Inf.* **193**, 105673 (2025).
24. Huang, T. Y., Hsieh, P. H. & Chang, Y. C. Performance comparison of junior residents and ChatGPT in the objective structured clinical examination (OSCE) for medical history taking and Documentation of medical records: development and usability study. *JMIR Med. Educ.* **10**, e59902 (2024).
25. Guo, Y. et al. Evaluating the accuracy, time and cost of GPT-4 and GPT-4o in liver disease diagnoses using cases from what is your diagnosis. *J. Hepatol.* **82** (1), e15–e17 (2025).
26. Gupta, G. K., Singh, A., Manikandan, S. V. & Ehtesham, A. *Digital diagnostics: the potential of large Language models in recognizing symptoms of common illnesses. Ai*, **6**(1). (2025).
27. Gunay, S., Ozturk, A. & Yigit, Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists. *Am. J. Emerg. Med.* **84**, 68–73 (2024).
28. Wang, J. et al. Adapting ChatGPT for color blindness in medical education. *Ann. Biomed. Eng.* **53** (1), 5–8 (2025).
29. Wang, G. et al. *Do Advanced Language Models Eliminate the Need for Prompt Engineering in Software Engineering?* arXiv:2411.02093 (2024). https://doi.org/10.48550/arXiv.2411.02093
30. Nori, H. et al. *From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond*. arXiv:2411.03590 (2024). https://doi.org/10.48550/arXiv.2411.03590
31. Brooks, N. *Is it possible to have pneumonia without a fever?* [cited 2025 Feb 12]; (2023). Available from: https://www.medicalnewstoday.com/articles/can-you-have-pneumonia-without-a-fever
32. Zelin, C. et al. Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT. *J. Biomed. Inf.* **157**, 104702 (2024).
33. Wang, C. et al. Ethical considerations of using ChatGPT in health care. *J. Med. Internet Res.* **25**, e48009 (2023).
34. Haltaufderheide, J. & Ranisch, R. *The ethics of ChatGPT in medicine and healthcare: a systematic review on large Language models (LLMs). Npj Digit. Med.*, **7**(1). (2024).
35. Tzelves, L. et al. ChatGPT in clinical medicine, urology and academia: A review. *Arch. Esp. Urol.* **77** (7), 708–717 (2024).
36. Kalodanis, K., Rizomiliotis, P. & Anagnostopoulos, D. European artificial intelligence act: an AI security approach. *Inform. Comput. Secur.* **32** (3), 265–281 (2024).
37. Triantafyllopoulos, L. et al. *Evaluating the interactions of medical Doctors with chatbots based on large Language models: insights from a nationwide study in the Greek healthcare sector using ChatGPT. Comput. Hum. Behav.*, 161. (2024).
38. Manolitsis, I. et al. Training ChatGPT models in assisting urologists in daily practice. *Stud. Health Technol. Inf.* **305**, 576–579 (2023).

Acknowledgements

NIH-NIGMS grants P20 GM103434, U54 GM-104942, and NSF 2125872 (GH). NIH-NLM grant No. R01LM013438 and NIDDK grant No. T32 DK137525 (LL). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The writing was polished by ChatGPT. We thank James Boulden, MD, MBA, from Reynolds Memorial and Wheeling Hospital for proof-reading the manuscript.

Author contributions

G.H.: conceptualization, formal analysis, and writing - original draft; JW: formal analysis, Writing - Review & Editing; K.S.: formal analysis, Writing - Review & Editing; L.L.: formal analysis, Writing - Review & Editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-95233-1.

Correspondence and requests for materials should be addressed to G.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.