

Zero-Shot Extraction of Structured Clinical Data from Free-Text CT Reports Using Foundational Large Language Models

Amulya Gupta, MBBS^{1,2*}

Alex Hongslo, MS-3^{1*}

Quynh Nguyen, MS-3¹

Jake Caldwell, MS-3¹

Ben Choi, MS-3¹

Christopher J. Harvey, B.S.^{1,2}

Jeffrey Thompson, Ph.D.^{1,3}

Diego Mazzotti, Ph.D.^{1,4}

Zijun Yao, Ph.D.⁵

Amit Noheria, MBBS, S.M.¹

¹The University of Kansas School of Medicine-Kansas City, Kansas City, KS

²Department of Cardiovascular Medicine

³Department of Biostatistics and Data Science

⁴Department of Medical Informatics

⁵Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS, USA

**Both authors contributed equally to this manuscript.*



MORE ONLINE
journals.ku.edu/kjm

Corresponding Author:

Amit Noheria, MBBS, S.M.

3901 Rainbow Blvd., MS 3901

Kansas City, KS 66160

Phone: (913) 588-9600

Email: noheriaa@gmail.com

ABSTRACT

Introduction. Large language models (LLMs) are being explored for multiple applications in medical research, including medical text classification. Authors of the study evaluated the performance of three off-the-shelf LLMs in classifying unstructured free-text CT pulmonary angiography reports for pulmonary embolism (PE)-related diagnostic labels.

Methods. Authors assessed human-labeled CT reports using three LLMs (ChatGPT-4o, Llama 3.3 70B, and Llama 3.1 8B) with zero-shot simple and chain-of-thought (CoT) prompts. Classification tasks included acute PE (N = 1,025) and, among reports with PE (n = 660), bilateral PE and large PE.

Results. LLMs demonstrated high classification performance for acute PE and bilateral PE but lower performance for large PE. With simple prompts, the highest F1 scores were 97.0% for acute PE (Llama 3.1 8B; sensitivity, 95.6%; specificity, 97.3%), 92.9% for bilateral PE (ChatGPT-4o; sensitivity, 98.0%; specificity, 88.6%), and 72.8% for large PE (ChatGPT-4o; sensitivity, 82.9%; specificity, 82.2%). CoT prompting did not meaningfully improve acute PE classification but provided slight improvements in subclassification performance for bilateral and large PE.

Conclusions. Off-the-shelf LLMs can classify free-text CT reports using simple prompt engineering and may facilitate automated data extraction for medical research following appropriate evaluation of classification performance.

INTRODUCTION

In clinical research, extracting structured data from unstructured medical reports is a potential application of large language models (LLMs). Historically, various natural language processing (NLP) systems have been considered the standard for automated information extraction from free-text documentation.¹ Although these systems computationally are efficient, they lack flexibility and often require substantial

task-specific customization and preprocessing.² In contrast, LLMs offer greater semantic flexibility and contextual understanding; however, they computationally are intensive, prone to hallucinations, and may not consistently follow prompt instructions.³

We therefore investigated the utility of readily available, off-the-shelf LLMs for extracting pulmonary embolism (PE) labels from free-text computed tomography pulmonary angiography (CTPA) reports using simple zero-shot and chain-of-thought (CoT) prompting. In simple zero-shot prompting, the LLM directly is prompted to provide a yes/no classification without worked input-output examples. In CoT prompting, the LLM is prompted to first generate a description of the report and then provide a yes/no classification, also without worked examples. Previous studies have evaluated LLMs for extracting structured information from free-text clinical notes.⁴⁻⁶ The aim of our study was to assess the reproducibility of these findings and directly compare simple zero-shot and CoT prompting strategies.

METHODS

This study was approved by the Institutional Review Board (IRB) of the University of Kansas Medical Center and did not require informed consent because of its retrospective design. The manuscript was prepared in accordance with the CHART (Chatbot Assessment Reporting Tool) statement for reporting evaluations of artificial intelligence chatbots, as recommended by the EQUATOR Network.⁷ We identified CTPA reports from the University of Kansas Health System from 2010 to 2022 using a natural language search for PE-related terms in Illuminate[®] Insight[™] (Softtek Illuminate Inc., Overland Park, KS), a clinical search engine, as part of a larger project evaluating PE diagnosis and outcomes. Patients with eligible CTPA reports were randomly assigned to four medical students for chart review using a REDCap[®] form. Each chart was reviewed by only one student, and the original CTPA report text was used for adjudication.

Anonymized CTPA report text was uploaded to a HIPAA-compliant, cloud-based institutional research platform built on Databricks[®] (Databricks, Inc., San Francisco, CA). Three LLM endpoints: ChatGPT-4o (released May 2024; closed-source, developed by OpenAI PBC, San Francisco, CA), Llama 3.3 (70 billion parameters; released De-

cember 2024; open-source, developed by Meta Platforms, Inc., Menlo Park, CA), and Llama 3.1 (8 billion parameters; released July 2024; open-source, developed by Meta Platforms, Inc., Menlo Park, CA) were queried independently on October 8, 2025, in Kansas City, USA, using the same prompts. The primary classification task was acute PE (yes/no), based on the radiologist's impression (Table 1). Among reports classified as positive for acute PE, additional sub-classification tasks assessed bilateral PE (defined as acute PE findings in both lungs, whether explicitly described or implicitly indicated) and large PE (defined as lobar involvement or greater). LLM outputs were post-processed using simple string-matching rules to extract binary labels.

CoT prompting was evaluated using the same three LLM endpoints. The prompts instructed the LLM to provide an explanation of the PE-related findings in the CTPA report followed by a binary conclusion. The output from this initial query was provided to a second LLM query, which was instructed to extract a single-word binary outcome from the preceding output.

For each classification task, we constructed voting ensemble models by aggregating the binary outputs from all three LLMs separately for simple and CoT prompting. Ensemble performance was evaluated at different voting thresholds based on the number of LLMs producing a positive classification.

We also evaluated the performance of the LLMs against a rule-based natural language processing (NLP) approach using regular-expression (Regex) string matching (detailed in the Supplementary Material, available online at journals.ku.edu/kjm). The matching rules were predefined based on domain expertise and were not iteratively refined after initial specification to avoid overfitting and ensure a fair comparison with the zero-shot LLM approaches.

LLM performance was evaluated using sensitivity, specificity, precision, accuracy, and F1 score, with human-labeled data serving as the reference standard. All analyses were performed using R[®] (version 4.4.1; R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Patient charts for 1,106 CTPA reports were manually reviewed for acute PE. Of these, 81 were excluded because of equivocal or undetermined findings, leaving 1,025 reports

Table 1. Prompts provided to the large language model for each report.

Task	Simple Prompt	Chain-of-Thought Prompt
Acute PE (Yes/No)	Evaluate the following CT scan report to determine if the patient has acute pulmonary embolism or not. This report is from a single patient only. Based on the contents of the clinical note, answer “yes” or “no” whether the report strongly suggest acute pulmonary embolism. Make sure to only return a single word answer, “yes” or “no” without any explanations, words, symbols. If the report is suggestive of chronic pulmonary embolism without any new acute PE, then conclude as “no”.	Evaluate the following CT scan report to determine if the patient has acute pulmonary embolism or not. This report is from a single patient only. Based on the contents of the clinical note, give a clear explanation about the description of pulmonary embolism related findings. Then conclude with a “yes” or “no” answer to whether acute pulmonary embolism is present or not. If the report is suggestive of chronic pulmonary embolism without any new acute PE, then conclude as “no”.
Bilateral PE (Yes/No)	Evaluate the following CT scan report of a patient with acute pulmonary embolism to determine whether the patient’s report suggests bilateral PE i.e. acute PE findings are present in both the lungs. This report is from a single patient only. Based on the contents of the clinical note, answer “yes” or “no” whether the report strongly suggest that acute PE is bilateral in this patient. Make sure to only return a single word answer, “yes” or “no” without any explanations, words, symbols. Do not include findings related to chronic PE in making this decision.	Evaluate the following CT scan report of a patient with acute pulmonary embolism to determine whether the patient’s report suggests bilateral PE i.e. acute PE findings are present in both the lungs. This report is from a single patient only. Based on the contents of the clinical note, give a clear explanation about the description of pulmonary embolism related findings. Then conclude with a “yes” or “no” answer to whether acute PE findings are present in both lungs or not. Do not include findings related to chronic PE in making this decision.
Extent of PE (Large/Small)	Evaluate the following CT scan report of a patient with acute pulmonary embolism to determine whether the patient’s report suggests large PE defined as affecting the lobar arteries, right or left pulmonary arteries, or being a saddle PE. The embolism should be classified as small if it is segmental or sub-segmental. If the affected arteries are not provided but the report categorically mentions it to be small, then answer it to be small. In case findings are suggestive of both small and large PE, answer it to be large. Based on the contents of the clinical note, answer “large” or “small” for acute PE in this patient. Make sure to only return a single word answer, “large” or “small” without any explanations, words, symbols. Do not include findings related to chronic PE in making this decision.	Evaluate the following CT scan report of a patient with acute pulmonary embolism to determine whether the patient’s report suggests large PE defined as affecting the lobar arteries, right or left pulmonary arteries, or being a saddle PE. The embolism should be classified as small if it is segmental or sub-segmental. If the affected arteries are not provided but the report categorically mentions it to be small, then answer it to be small. In case findings are suggestive of both small and large PE, answer it to be large. Based on the contents of the clinical note, give a clear explanation about the description of pulmonary embolism related findings. Then conclude with a “large” or “small” answer for acute PE. Do not include findings related to chronic PE in making this decision.

for analysis. Acute PE was present in 660 (64.4%) reports; among the 642 reports with non-missing bilateral labels, 300 (46.7%) were bilateral and among the 660 reports with acute PE, 187 were classified as large PE. Manually extracted yes/no labels for bilateral PE were missing for 18 (2.7%) patients, who were excluded from the bilateral PE analysis.

Acute PE (Yes/No)

As shown in Table 2, with simple prompting, Llama 3.1 8B achieved the numerically highest F1 score (97.0%), followed by ChatGPT-4o (96.3%) and Llama 3.3 70B (95.5%) for acute PE classification. With CoT prompting, performance declined across all three models (F1 scores: ChatGPT-4o,

95.6%; Llama 3.3 70B, 94.7%; Llama 3.1 8B, 88.8%). The rule-based NLP approach had an F1 score of 75.8%.

Bilateral PE (Yes/No)

With simple prompting, ChatGPT-4o achieved the numerically highest F1 score (92.9%), followed by Llama 3.3 70B (90.3%) and Llama 3.1 8B (84.5%). With CoT prompting, ChatGPT-4o and Llama 3.3 70B showed slight improvements in performance (F1 scores, 93.6% and 93.8%, respectively), whereas Llama 3.1 8B showed no change (84.5%). The rule-based NLP approach had an F1 score of 66.7%.

Table 2. Performance of large language models in CTPA report classification tasks for pulmonary embolism labels.

Classification Task and Model	Simple Prompting						Chain-of-Thought Prompting					
	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)		Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)	
Acute PE (N = 1025)												
Rule-based NLP	78.0	49.9	73.8	68.0	75.8		-	-	-	-	-	
ChatGPT-4o	93.8	98.1	98.9	95.3	96.3		92.4	98.4	99.0	94.5	95.6	
Llama 3.3 70B	92.1	98.6	99.2	94.4	95.5		90.8	98.4	99.0	93.5	94.7	
Llama 3.1 8B	95.6	97.3	98.4	96.2	97.0		91.1	74.5	86.6	85.2	88.8	
Bilateral Acute PE (n = 642)												
Rule-based NLP	81.7	44.4	56.3	61.8	66.7		-	-	-	-	-	
ChatGPT-4o	98.0	88.6	88.3	93.0	92.9		97.0	90.9	90.4	93.8	93.6	
Llama 3.3 70B	99.0	82.2	83.0	90.0	90.3		98.0	90.4	89.9	93.9	93.8	
Llama 3.1 8B	98.0	70.2	74.2	83.2	84.5		98.7	69.3	73.8	83.0	84.5	
Large PE (n = 660)												
Rule-based NLP	63.1	76.3	51.3	72.6	56.6		-	-	-	-	-	
ChatGPT-4o	82.9	82.2	64.9	82.4	72.8		84.0	83.1	66.2	83.3	74.1	
Llama 3.3 70B	80.2	82.9	64.9	82.1	71.8		79.1	84.4	66.7	82.9	72.4	
Llama 3.1 8B	87.2	62.4	47.8	69.4	61.7		82.4	68.5	50.8	72.4	62.9	

Extent of PE (Large/Small)

Classification performance was lowest for large PE compared with the other two classification tasks. With simple prompting, ChatGPT-4o had the highest F1 score (72.8%), followed by Llama 3.3 70B (71.8%) and Llama 3.1 8B (61.7%). With CoT prompting, there were slight numerical improvements in performance across all three models (F1 scores: ChatGPT-4o, 74.1%; Llama 3.3 70B, 72.4%; Llama 3.1 8B, 62.9%). The rule-based NLP approach had an F1 score of 56.6%.

As shown in Supplementary Tables 1 and 2 (available online at journals.ku.edu/kjm), combining the outputs of the three LLMs in voting ensemble models resulted in slight absolute improvements in performance. For the simple-prompt voting ensemble, the highest F1 scores were 97.6% for acute PE (≥ 1 vote), 93.3% for bilateral PE (≥ 3 votes), and 73.2% for large PE (≥ 3 votes). For the CoT-prompt voting ensemble, the highest F1 scores were 95.3% for acute PE (≥ 2 vote), 94.5% for bilateral PE (≥ 2 votes), and 72.7% for large PE (≥ 3 votes).

DISCUSSION

We evaluated the performance of three publicly available, off-the-shelf LLMs for extracting PE-related diagnostic labels from free-text CTPA reports. The key findings of our study are as follows. First, all tested models, including the smaller and computationally efficient Llama 3.1 8B, demonstrated high classification performance for the relatively straightforward task of identifying acute PE. Second, performance declined across models for more nuanced classification tasks, including identification of bilateral PE and large PE. Although Llama 3.1 8B maintained relatively strong performance for bilateral PE, it had the lowest F1 scores among the three models for both subclassification tasks, suggesting that larger or better-tuned models may be more reliable for more nuanced information extraction. Third, CoT prompting did not improve performance for the straightforward primary classification task but may provide modest improvements for more nuanced classifications. Fourth, voting ensemble models that aggregated binary outputs from multiple LLMs yielded only marginal improvements in overall performance; however, adjusting the

voting threshold provided a means of tuning the balance between sensitivity and specificity.

This analysis builds upon several recent studies evaluating LLMs for structuring free-text clinical data.⁸⁻¹² As reported in these studies, although LLMs relatively are easy to deploy and can achieve high classification performance, they may produce inconsistent outputs that are difficult to predict and systematically mitigate.¹³ LLMs are also prone to hallucinations, in which they generate information or labels that are unsupported by the input or assign classifications when the available information is ambiguous.¹⁴

LLMs are based on transformer architectures that use contextual representations to predict subsequent tokens in a sequence.¹⁵ Unlike conventional rule-based NLP approaches, LLMs do not rely solely on explicitly defined terms or patterns to assign classifications. This flexibility may contribute to their ability to recognize contextual relationships in free-text reports but also makes their outputs less transparent and more probabilistic. Moreover, in most LLM endpoints, the prompt instructions and input text, such as clinical reports, are provided as part of a single input sequence. This may make LLM-based systems sensitive to irregularities in free-text input, including unexpected formatting, special characters, or potentially adversarial instructions embedded within the input.

Nonetheless, our findings demonstrate good empirical performance of zero-shot LLMs for free-text classification, particularly for relatively straightforward tasks such as acute PE diagnosis. Based on these observations, we present a general framework for binary label classification using LLMs in Figure 1.

CoT prompting is a technique in which the model is prompted to generate intermediate reasoning steps before providing a final answer, which may improve classification performance and provide additional information for reviewing the basis of a classification.¹⁶ In our dataset, performance for the primary classification of acute PE was lower with CoT prompting than with simple prompting. The decline was most pronounced for the smallest LLM, Llama 3.1 8B, suggesting that smaller models may be less suitable for CoT prompting in this setting. In contrast, CoT prompting produced modest improvements in performance for the subclassification tasks of bilateral and large PE.

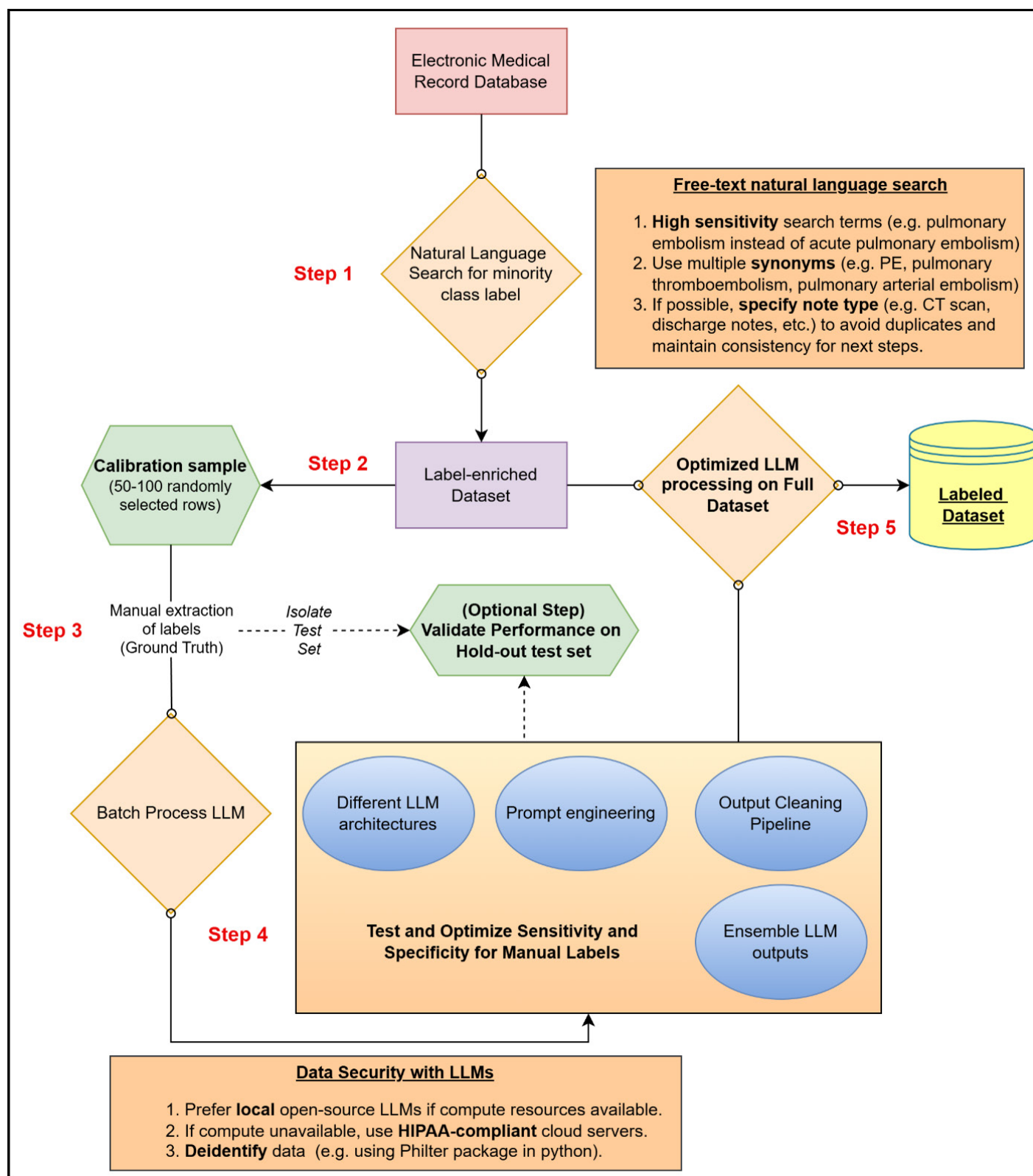


Figure 1. A general outline for binary label classification of free-text clinical reports using a simple large language model and natural language search framework.

Alternatives to LLMs include rule-based NLP approaches. However, these approaches can be brittle to minor variations in language and may therefore have limited sensitivity or specificity.^{17,18} Machine-learning and hybrid NLP approaches are other alternatives for text classification and can achieve high classification performance but generally require substantial numbers of training samples and extensive testing before deployment.^{19,20} Fine-tuning LLMs can improve classification performance but is resource-intensive, requires labeled datasets, and produces task-specific models that may be less practical for smaller-scale research projects.²¹

Limitations

First, we used label-enriched reports that were prefiltered using a natural language search for PE-related terms. This sampling approach may have resulted in a study population that differs from the broader population of CTPA reports encountered in clinical practice and may have affected observed classification performance. Second, each case was manually labeled by a single reviewer from a group of four medical students without independent cross-verification. In addition, reviewers may have considered broader clinical information from the patient charts beyond the CTPA reports when assigning labels, whereas the LLMs were provided only with the CTPA report text. This difference between the human labeling process and the LLM input may have introduced misclassification and affected estimates of model performance. Third, we did not evaluate the effects of varying LLM hyperparameters or prompt formulations on performance. Finally, larger studies involving more diverse clinical reports and classification tasks are needed to evaluate the generalizability of our findings.

CONCLUSIONS

Off-the-shelf LLMs effectively can classify free-text CTPA reports into diagnostic labels using simple zero-shot prompts. In this study, larger models, Llama 3.3 70B and ChatGPT-4o, generally demonstrated higher observed performance than the smaller Llama 3.1 8B model, particularly for more nuanced classification tasks. Further research is warranted to evaluate the utility of LLMs for medical text classification across diverse clinical datasets, prompting strategies, model architectures, and hybrid NLP approaches.

ARTICLE INFORMATION

Received April 22, 2026; Accepted for publication September 8, 2026; Published online October 16, 2026, *Kans J Med* 2026 Sep-Oct; 19:125-132. <https://doi.org/10.17161/kjm.vol19.25396>.

Disclosures: ChatGPT was used solely for grammatical and stylistic refinement after the initial draft was prepared by the authors. The authors reviewed and approved the final manuscript and take full responsibility for the accuracy, integrity, and attribution of its content. The authors report no relevant financial conflicts of interest.

REFERENCES

1. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J Am Med Inform Assoc*. 2016; 23(5):1007–15. Epub 2016 Feb 5. PMID: 26911811.
2. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. 2017; 24(5):986–991. PMID: 28419261.
3. Lu Z, Peng Y, Cohen T, Ghassemi M, Weng C, Tian S. Large language models in biomedicine and health: Current research landscape and future directions. *J Am Med Inform Assoc*. 2024; 31(9):1801–1811. PMID: 39169867.
4. Joseph E, Vallee P, Perennec T, et al. Development and assessment of a pipeline for extracting structured data from free-text medical reports using a large language model. *JCO Clin Cancer Info*. 2026; 10(1): e2500133. PMID: 41707099.
5. Garcia-Carmona AM, Prieto ML, Puertas E, Beunza JJ. Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study. *JMIR AI*. 2025; 4:e68776. PMID: 40608403.
6. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. *NPJ Digit Med*. 2024 Aug 24; 7(1):222. PMID: 39182008.
7. Huo B, Collins G, Chartash D, et al. Reporting guideline

- for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *BMJ Med.* 2025; 4(1):e001632. PMID: 40761518.
8. Le Guellec B, Lefèvre A, Geay C, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiol Artif Intell.* 2024; 6(4):e230364. PMID: 38717292.
 9. Abdullah A, Kim ST. Automated radiology report labeling in chest X-ray pathologies: Development and evaluation of a large language model framework. *JMIR Med Inform.* 2025; 13:e68618. PMID: 40153539.
 10. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. *Int J Med Inform.* 2024; 183:105321. PMID: 38157785.
 11. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit Med.* 2024; 7(1):106. PMID: 38693429.
 12. Burford KG, Itzkowitz NG, Ortega AG, Teitler JO, Rundle AG. Use of generative AI to identify helmet status among patients with micromobility-related injuries from unstructured clinical notes. *JAMA Netw Open.* 2024; 7(8):e2425981. PMID: 39136946.
 13. Shah SV. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Netw Open.* 2024; 7(8):e2425953. PMID: 39136951.
 14. Bijoy Das A, Ahmed S, Sakib SK. Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models. *arXiv e-prints.* 2025:arXiv:2504.19061. doi:10.48550/arXiv.2504.19061.
 15. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv e-prints.* 2017:arXiv:1706.03762. doi:10.48550/arXiv.1706.03762.
 16. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. *Medicina (Kaunas).* 2024; 60(1):148. PMID: 38256408.
 17. Rashedi S, Bukhari S, Krishnathan D, et al. Optimizing the accuracy of natural language processing tools for pulmonary embolism detection through integration with claims data: The PE-EHR+ study. *Thromb Haemost.* Epub 2026 Feb 9. PMID: 41605431.
 18. Somani SS, Honarvar H, Narula S, et al. Development of a machine learning model using electrocardiogram signals to improve acute pulmonary embolism screening. *Eur Heart J Digit Health.* 2021; 3(1):56-66. PMID: 35355847.
 19. Hassan SM, Mylvaganam R, Didebulidze T, et al. Leveraging hybrid natural language processing techniques for large-scale pulmonary embolism identification. *JACC Adv.* 2025; 4(11 Pt 2):101845. Epub 2025 May 21. PMID: 41350101.
 20. Chen MC, Ball RL, Yang L, et al. Deep learning to classify radiology free-text reports. *Radiology.* 2018; 286(3):845-852. Epub 2017 Nov 13. PMID: 29135365.
 21. Shankar SV, Dhingra LS, Aminorroaya A, et al. Automated transformation of unstructured cardiovascular diagnostic reports into structured datasets using sequentially deployed large language models. *Eur Heart J - Digit Health.* 2025; 6(4):783-796. PMID: 40703108.

Keywords: *large language models; artificial intelligence; natural language processing; diagnosis; electronic health records*