ORIGINAL RESEARCH **Open access**

Large Language Model with Retrieval-Augmented Generation and Chain-of-Thought Reasoning for Differential Diagnosis Generation from Emergency Department Triage Notes and Vital Signs

Isabella Garcia[1*], Diego Herrera[1]

Abstract

This article proposes a conceptual framework for a diagnostic support system in emergency departments that leverages large language models, retrieval-augmented generation, and chain-of-thought reasoning. By combining triage notes and vital signs, the system generates a ranked differential diagnosis list to assist clinicians without replacing their judgment. The framework includes components like a triage note encoder, a vital sign encoder, a retrieval module, and a diagnosis ranker, using evidence from clinical guidelines, curated references, and de-identified prior cases. The approach grounds the model in authoritative knowledge while ensuring transparency and explainability in the diagnostic process. However, prospective validation, integration into workflows, and clinician oversight are crucial before implementation to ensure safety and effectiveness.

Keywords Clinical decision support, Large language models, Retrieval-augmented generation, Chain-of-thought reasoning, Emergency department triage, Differential diagnosis

*Correspondence:
Isabella Garcia
isabella.garcia@gmail.com
[1] Department of Clinical AI Systems, University of Buenos Aires, Buenos Aires, Argentina

Introduction

Emergency department diagnosis begins under conditions of uncertainty, incomplete history, crowding, and time pressure, yet the earliest data elements may already contain clinically meaningful signals. Triage notes summarize chief complaint, symptom duration, mechanism, and contextual risk factors, while vital signs provide structured physiologic evidence of instability or occult deterioration. Recent work on large language model assessment in emergency settings has shown that acuity classification, admission prediction, discharge documentation, and patient-facing emergency care questions are increasingly tractable computational tasks, although none remove the need for clinical interpretation [1-5]. A framework for differential diagnosis generation from triage notes and vital signs therefore addresses a practical point in the emergency workflow where rapid hypothesis formation can shape early testing, prioritization, and escalation.

Large language models have demonstrated broad clinical reasoning capability across medical examinations, diagnostic vignettes, and complex clinical cases, suggesting that they can synthesize heterogeneous clinical information into plausible diagnostic hypotheses. Med-PaLM and related systems show that model scale and instruction tuning can encode clinically relevant knowledge, while GPT-4 evaluations suggest performance on difficult diagnostic challenges and medical records with delayed diagnoses [6-9]. However, the same generative capacity that enables flexible reasoning also creates risk when

models produce unsupported explanations, omit dangerous diagnoses, or express unwarranted certainty. Reviews of large language models in medicine emphasize that clinical usefulness depends not only on accuracy but also on safety, evaluation, transparency, and governance [10].

Retrieval-augmented generation offers one route to reduce unsupported generation by connecting model outputs to external clinical knowledge at inference time. In healthcare, retrieval can incorporate hospital protocols, clinical guidelines, de-identified prior cases, drug databases, and specialty references, thereby narrowing the model's evidence base to sources relevant to the current patient context [11-14]. Chain-of-thought prompting complements retrieval by encouraging explicit intermediate reasoning rather than a direct answer, which may improve diagnostic transparency and enable clinicians to inspect the logic behind a ranked differential [15-17]. The proposed framework combines these two mechanisms so that the model reasons from both patient-specific inputs and retrieved clinical evidence.

This article presents a conceptual architecture for large language model-based differential diagnosis generation using emergency department triage notes and vital signs as the primary inputs. It does not report experiments, performance claims, or fabricated validation results; instead, it defines system components, design principles, safety constraints, and evaluation pathways informed by recent peer-reviewed work on LLMs, RAG, CoT, emergency medicine, and diagnostic support. The manuscript first reviews ED triage, differential diagnosis, clinical LLM reasoning, and RAG-CoT methods, then describes an architecture for retrieval-grounded reasoning and structured diagnosis ranking [18-21]. The central thesis is that an ED-facing LLM should not operate as an unconstrained chatbot, but as a bounded, evidence-linked, clinician-supervised reasoning system.

Background

ED triage process

Emergency department triage is designed to rapidly stratify urgency using presenting complaint, brief history, observed distress, comorbid risk, and vital signs such as heart rate, blood pressure, respiratory rate, oxygen saturation, and temperature. Triage scales such as the Emergency Severity Index and other acuity systems impose structure on early assessment, but real-world triage still depends on concise narrative notes and contextual interpretation. Comparative evaluations of LLMs in emergency triage show growing interest in whether models can approximate acuity judgments, although specialist performance and institutional workflow remain important reference points [2, 21, 22]. For differential diagnosis generation, triage data should therefore be treated as an early, incomplete, and high-value snapshot rather than a full clinical encounter.

Differential diagnosis generation

Differential diagnosis generation is an iterative cognitive process in which clinicians collect early data, generate candidate hypotheses, weigh supporting and refuting evidence, and revise the list as new information arrives. Diagnostic errors can occur when clinicians anchor on an initial impression, close the search too early, underweight abnormal vital signs, or fail to consider rare but dangerous conditions. Studies of LLM diagnostic performance in complex cases, pediatric case studies, and difficult clinical scenarios suggest that models may broaden hypothesis generation, but they also require careful oversight when moving from vignette-based reasoning to active clinical decision support [7, 8, 23, 24]. In the emergency department, a useful LLM system should help surface plausible and "do-not-miss" diagnoses without displacing clinician accountability.

LLMs for clinical reasoning

Large language models such as GPT-4, Med-PaLM, and other instruction-tuned medical models have shown capacity for medical question answering, diagnostic reasoning, and synthesis of clinical narratives. Singhal *et al.* demonstrated that large language models can encode clinical knowledge, while subsequent clinical evaluations examined their behavior on diagnostic challenges, delayed diagnosis records, and physician-facing reasoning tasks [6, 9, 19]. Emergency medicine applications have also begun to test LLMs for triage assessment, admission prediction, discharge documentation, and response quality for patient questions [1, 3, 5, 25]. Nevertheless, most published results remain task-specific, and real-time ED diagnosis requires additional safeguards for latency, incomplete inputs, hallucination, and institutional accountability.

RAG and CoT

Retrieval-augmented generation refers to the process of retrieving relevant external documents before response generation, allowing the model to condition its answer on

specific passages rather than relying only on parametric memory. In clinical settings, this can include hospital guidelines, validated medical references, de-identified clinicopathologic cases, pharmacology databases, and local protocols that reflect institutional practice [11-13, 26]. Chain-of-thought prompting asks the model to expose or structure intermediate reasoning steps, and medical studies have examined its potential to improve reasoning quality in question answering, nephrology, laboratory medicine, and radiology interpretation [15-17, 27]. Combining RAG and CoT is therefore conceptually attractive because retrieval supplies evidence while structured reasoning organizes how that evidence is applied.

Framework Overview

High-level architecture

The proposed architecture begins with two primary inputs available at the earliest stage of emergency care: the free-text triage note and structured vital signs recorded at presentation. A triage note encoder converts chief complaint, symptom descriptors, temporal patterns, comorbid risk factors, medication cues, mechanism of injury, and contextual details into a semantic representation that preserves clinically meaningful relationships. In parallel, a vital sign encoder transforms heart rate, blood pressure, respiratory rate, oxygen saturation, temperature, pain score, and mental-status indicators into clinically interpretable abnormality features, allowing physiologic instability to shape retrieval and reasoning rather than appearing as isolated numeric values. These combined representations form the query for a retrieval system that identifies relevant guidelines, similar de-identified cases, and clinical reference passages before the LLM performs structured reasoning and produces a ranked differential diagnosis list [11, 14, 18]. The final output should include candidate diagnoses, supporting and opposing evidence, urgency flags, missing information, and source attribution rather than a single unsupported answer, thereby positioning the system as a reasoning aid for clinicians rather than an autonomous diagnostic agent.

Figure 1 presents the proposed retrieval-grounded and reasoning-aware architecture for transforming emergency department triage notes and vital signs into clinician-supervised ranked differential diagnosis support.

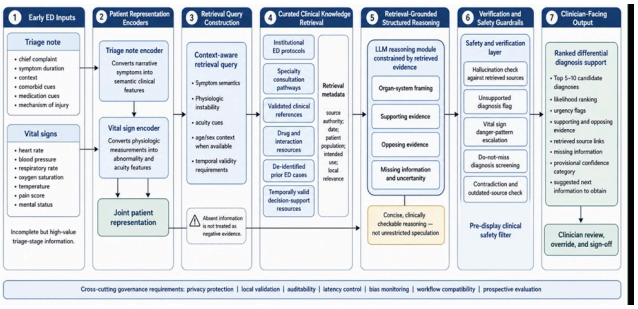

Figure 1. Retrieval-Grounded and Reasoning-Aware LLM Architecture for Emergency Department Differential Diagnosis Generation

Core assumptions

This framework assumes that the hospital has digitized triage documentation, structured vital signs, and access to a curated clinical knowledge base compatible with privacy, security, and institutional governance requirements. It also assumes that the system is used at the point of care, where inference time must be short enough to fit the emergency department workflow and where retrieved passages must be concise enough for rapid clinician review. Because ED triage data are inherently incomplete, the model must be designed to operate under uncertainty, distinguish absent information from negative findings, and avoid overstating diagnostic confidence when history, examination, laboratory data, or imaging are not yet available. Prior emergency department LLM studies on triage, admission prediction, and documentation suggest that early clinical text can support computational decision-support tasks, but they also show that performance should be assessed in relation to local practice, patient population, documentation style, and prospective clinician use [1-4]. The system is therefore designed as an assistive layer that augments early hypothesis generation, prioritization, and safety checking rather than functioning as an autonomous diagnostic authority.

Design principles

The first design principle is explainability: the model should make clear which triage details, vital sign abnormalities, retrieved passages, and reasoning steps influenced each candidate diagnosis. The second is provenance: retrieved evidence should be linked to source documents so that clinicians can distinguish guideline-grounded reasoning from model inference, local policy from general medical knowledge, and current recommendations from potentially outdated material. The third is safety: emergent diagnoses

must be flagged even when unlikely, uncertainty must be visible, and high-stakes recommendations must require clinician sign-off before being acted upon [5, 10, 19]. A fourth principle is workflow compatibility, because an ED-facing system must produce concise, scannable, and actionable output without increasing documentation burden or distracting clinicians during time-sensitive care. Together, these principles align with the broader movement from general-purpose medical chatbots toward bounded clinical systems with auditable inputs, outputs, retrieval traces, and reasoning pathways.

Table 1 clarifies how each component of the proposed framework contributes to diagnostic support while addressing a distinct safety, interpretability, or workflow failure mode.

Table 1. Design Logic of the Proposed ED Differential Diagnosis Framework

Framework layer	Design function	Clinical rationale	Failure r addres
Triage note encoder	Converts brief narrative documentation into clinically meaningful symptom, timing, context, and risk features	ED reasoning often begins from incomplete narrative information before diagnostic testing is available	Loss of cor clues; c reliance isolated ke
Vital sign encoder	Converts physiologic measurements into acuity and abnormality features	Vital signs may reveal instability, occult deterioration, or “do-not-miss” patterns	Treating ab physiolo secondary genera
Joint patient representation	Combines narrative and structured physiologic evidence	Differential diagnosis requires integration of symptoms with physiologic state	Fragme reason between te numeric
Retrieval module	Retrieves relevant institutional, clinical, and case-based evidence	RAG reduces reliance on parametric memory and supports source-grounded reasoning	Hallucinate outdat recommen unsupp diagnostic
Structured reasoning module	Organizes evidence into support, opposition, uncertainty, and missing information	Clinicians need inspectable reasoning rather than opaque diagnostic lists	Fluent clinically u explana
Diagnosis ranker	Produces a prioritized differential diagnosis list	ED decision-making requires both likely diagnoses and urgent exclusions	Premat closure; or of rare danger diagno
Verification layer	Checks output against retrieval evidence, internal consistency, and safety constraints	High-stakes clinical output must be reviewed before display	Unsupp diagnos contradic unsafe om
Clinician interface	Presents ranked output, evidence links, uncertainty, and override options	The system must augment rather than replace emergency clinician judgment	Automatio uncle accounta

Rag for Clinical Context

Knowledge base construction

The RAG component requires a curated knowledge base that reflects both general medical evidence and local emergency department practice. Candidate sources include institutional protocols, specialty consultation pathways, drug interaction databases, de-identified prior ED cases, radiology and laboratory decision aids, validated clinical references, and symptom-specific emergency pathways, with each document labeled by topic, date, source authority, patient population, and intended use. The knowledge base should be organized so that high-acuity and symptom-oriented content is readily retrievable, because ED differential diagnosis often begins from presentations such as chest pain, shortness of breath, altered mental status, fever, syncope, abdominal pain, trauma, or neurologic deficit rather than from confirmed disease categories. Healthcare RAG studies and reviews emphasize that retrieval quality depends on corpus curation, chunking strategy, indexing method, metadata design, update frequency, and governance over what content is allowed to ground clinical recommendations [12-14, 26]. For ED differential diagnosis, the knowledge base should therefore prioritize time-sensitive, high-acuity, locally relevant, and clinically validated materials rather than broad textbook content alone.

Retrieval strategy

At inference time, the system should encode the triage note and vital signs as a joint query, retrieve the top relevant passages, and rerank them according to symptom match, physiologic abnormality, acuity, patient age, comorbidity context, and temporal validity. The retrieval query should preserve both semantic and structured information, so that terms such as “chest pressure,” “sudden onset,” “hypotension,” “tachycardia,” “fever,” or “low oxygen saturation” influence which guidelines and prior cases are surfaced. For example, a patient with chest pain, hypotension, tachycardia, and hypoxemia should retrieve not only common causes such as acute coronary syndrome and pulmonary embolism, but also “do-not-miss” entities such as aortic dissection, tension pneumothorax, sepsis, and massive hemorrhage when clinically plausible. Prior RAG work in clinical decision support shows that retrieval can improve the factual grounding of LLM outputs, but it also introduces risks if retrieved passages are outdated, poorly matched, incomplete, or over-weighted by the generator [11, 12, 14]. The framework therefore treats retrieval as evidence selection rather than automatic truth, requiring the model to use retrieved material critically and to expose uncertainty when the retrieved evidence is weak or conflicting.

Chain-of-Thought Reasoning

Prompt engineering

The prompt should define the LLM’s role as an emergency department diagnostic support assistant that generates differential diagnoses from incomplete early information. The user message should include the triage note, vital signs, patient age and sex when available, and the retrieved passages, followed by explicit instructions to reason systematically across organ systems, acuity, supporting evidence, refuting evidence, and immediate danger. The prompt should also instruct the model to identify missing information, distinguish abnormal vital signs from normal or borderline values, and avoid assuming that unmentioned findings are absent. Studies of chain-of-thought reasoning in medical question answering, nephrology, laboratory medicine, and radiology suggest that structured reasoning can improve interpretability and may support more complete diagnostic analysis, although it can still produce plausible but incorrect logic [15-17, 27]. The prompt should therefore ask for concise, clinically checkable reasoning rather than unrestricted narrative speculation, while making clear that the output is provisional and requires clinician review.

Structured output

The model output should follow a stable structure that begins by identifying likely organ systems, then lists candidate conditions, interprets vital sign abnormalities, highlights emergent exclusions, and produces a ranked differential diagnosis. Each diagnosis should include supporting evidence from the triage note, relevant vital sign patterns, retrieved source support, opposing evidence, missing information, urgency level, and a confidence category that remains explicitly provisional. The ranked list should distinguish high-probability diagnoses from high-risk diagnoses, because an immediately dangerous condition may require urgent evaluation even when it is not the most statistically likely explanation for the presentation. Diagnostic LLM studies using radiology quizzes, complex clinical cases, and familial inflammatory disorders show that model reasoning can be useful when organized around evidence, but the output must remain reviewable and contestable by clinicians [28-30]. A structured format also enables downstream evaluation of whether reasoning steps

are logical, clinically plausible, aligned with retrieved evidence, and responsive to the constraints of emergency department decision-making.

Differential Diagnosis Generation

Ranked list format

The differential diagnosis output should present the top five to ten diagnoses in ranked order, with each entry justified by patient-specific findings and retrieved clinical evidence. For each diagnosis, the model should state why the triage note and vital signs support consideration of that condition, what information weakens it, and what immediate tests or bedside assessments would usually clarify the probability. Emergency-facing LLM studies show that models can assist with acuity estimation, admission prediction, and diagnostic generation, but the ranked list must explicitly preserve “do-not-miss” conditions even when their estimated likelihood is lower [1, 3, 20]. A chest pain case, for instance, should not only rank common causes but also flag aortic dissection, pulmonary embolism, tension pneumothorax, sepsis, and acute coronary syndrome when the clinical pattern warrants urgent exclusion.

Confidence calibration

Confidence should be represented as provisional clinical uncertainty rather than as a definitive probability of disease. A practical system could combine verbal confidence categories, retrieval agreement, consistency across repeated reasoning passes, and model-estimated uncertainty to distinguish high-support diagnoses from speculative alternatives. Randomized and comparative evaluations of LLM diagnostic reasoning suggest that model influence can affect clinician reasoning, making calibration and uncertainty communication essential for safe use [19, 31, 32]. The framework therefore treats confidence as an aid to prioritization, not as a replacement for physician judgment, examination, laboratory testing, imaging, or reassessment.

Interpretation and Explanation

Explanation user interface

The user interface should show each diagnosis with linked retrieved evidence, concise reasoning steps, abnormal vital sign interpretation, and missing information that would change the ranking. Instead of exposing a long free-form rationale, the interface should separate evidence into clinically meaningful fields such as “supports,” “argues against,” “urgent exclusions,” and “next information needed.” Prior work on LLMs for emergency handoff notes, discharge documentation, and patient-facing emergency questions shows that clarity, brevity, and clinical relevance are central to safe communication in ED workflows [4, 5, 25]. Explanation should therefore be designed for rapid review by clinicians who are managing multiple patients simultaneously.

Clinician override

Clinician override should be a first-class system function rather than an afterthought. The clinician should be able to mark a diagnosis as irrelevant, add a missing diagnosis, flag incorrect reasoning, identify unsafe retrieval, or indicate that a vital sign abnormality was artifactually measured. Feedback could later guide retrieval weighting, prompt revision, local governance review, or supervised model improvement, but it should not automatically update the system without validation. Because LLM studies in diagnosis show both promising synthesis and meaningful risk of misleading outputs, clinician correction must remain visible, auditable, and institutionally governed [7, 10, 23, 24].

Safety and Hallucination Mitigation

Hard guardrails

Hard guardrails should prevent the system from producing overconfident diagnostic recommendations when critical information is absent, contradictory, or physiologically unstable. If vital signs indicate shock, hypoxemia, altered mental status, or other high-risk patterns, the model should prioritize immediate escalation language and “do-not-miss” diagnoses rather than low-acuity explanations. Reviews and clinical studies of medical LLMs repeatedly emphasize that safety depends on transparency, bounded use, uncertainty messaging, and prevention of unsupported claims [6, 10, 13, 14]. The system should therefore refuse unsupported specificity, label missing data clearly, and avoid suggesting that the ranked list is exhaustive.

Post-hoc verification

Post-hoc verification should compare the generated differential against retrieved passages, medication facts, symptom definitions, vital sign thresholds, and internal consistency checks before display. This verification layer can flag unsupported diagnoses, conflated diseases, irrelevant retrieved passages, contradictory reasoning, or unsafe omissions of emergent conditions. RAG and CoT studies suggest that retrieval and reasoning can improve clinical usefulness, but neither mechanism guarantees truthfulness when the source evidence is mismatched or the reasoning chain is superficially plausible [11, 12, 15, 16]. For high-stakes recommendations, the output should be treated as a draft clinical reasoning artifact requiring physician review and sign-off.

Evaluation Strategy

Diagnostic accuracy

Diagnostic accuracy should be evaluated by comparing the model's top-one, top-three, and top-five differential diagnoses against clinician-generated differentials, final ED diagnoses, hospital discharge diagnoses, or adjudicated expert panels. Metrics should include top-k accuracy, precision, recall for critical diagnoses, and performance stratified by chief complaint, acuity, age group, and abnormal vital sign pattern. Prior diagnostic studies involving GPT-4, medical case challenges, radiology quizzes, and complex clinical records demonstrate the feasibility of benchmarking diagnostic reasoning, but ED deployment requires evaluation on early, incomplete triage-stage information rather than polished case vignettes [8, 9, 18, 28, 29]. The most important accuracy target is not merely naming the final diagnosis, but safely prioritizing dangerous possibilities early enough to influence care.

Table 2 defines the evaluation domains needed to move the proposed system from conceptual architecture toward safe emergency department simulation and prospective validation.

Table 2. Evaluation Matrix for Safe Deployment of Retrieval-Grounded LLM Differential Diagnosis Support

Evaluation domain	Core question	Suggested assessment approach	Minimum reporting requireme
Diagnostic prioritization	Does the system place plausible and dangerous diagnoses appropriately in the ranked list?	Top-1, top-3, and top-5 comparison against clinician differential, final ED diagnosis, discharge diagnosis, or expert adjudication	Performan by chief complaint acuity grou age group and abnorma vital sign pattern
Critical diagnosis recall	Does the system preserve "do-not-miss" conditions even when likelihood is uncertain?	Condition-specific recall for sepsis, acute coronary syndrome, pulmonary embolism, stroke, aortic dissection, shock, and hypoxemia-related emergencies	Separate reporting f high-risk diagnose rather tha aggregate accuracy alone
Retrieval relevance	Are retrieved passages clinically relevant, current, and source-appropriate?	Blinded clinician rating of retrieved evidence match, source authority, temporal validity, and local applicability	Retrieval precision outdated source frequency and irrelevant source frequency
Reasoning quality	Is the diagnostic logic clinically valid and aligned with available evidence?	Emergency physician scoring of support, opposition, missing information, uncertainty, and vital sign interpretation	Independe reasoning quality sco separate from diagnosis accuracy

Hallucination control	Are generated claims supported by retrieved sources and patient inputs?	Post-hoc source attribution audit and unsupported-claim detection	Rate of unsupporte diagnosis unsupporte rationale and unsupporte manageme suggestio
Calibration and uncertainty	Does the system communicate provisional confidence appropriately?	Agreement between confidence category, retrieval support, repeated-pass consistency, and clinician judgment	Calibratio plots or category-level erro rates wher feasible
Workflow usability	Can clinicians review the output quickly during ED work?	Simulation-based usability testing, time-to-review, cognitive load rating, and clinician override frequency	Median latency, review tim override ra and perceived usefulnes
Prospective safety	Does the system remain safe under real triage conditions?	Silent-mode prospective simulation before active deployment	Error taxonomy near-miss analysis, subgroup performanc and escalatio failures
Bias and subgroup robustness	Does performance differ across documentation styles, language groups, age groups, sex, acuity, or comorbidity profiles?	Stratified evaluation across demographic, clinical, and documentation subgroups	Subgroup specific diagnosti and reasoning metrics
Governance and auditability	Can outputs, sources, overrides, and failures be reviewed institutionally?	Audit log review, governance committee evaluation, and safety incident tracking	Traceable record of inputs, retrieved sources, output, clinician response and overri

Reasoning quality

Reasoning quality should be assessed independently from final diagnosis accuracy because a correct diagnosis can be reached through incomplete, biased, or clinically unsafe logic. Blinded emergency physicians could rate model reasoning for logical completeness, relevance to the triage note, appropriate use of vital signs, evidence alignment, recognition of uncertainty, and inclusion of urgent exclusions. Chain-of-thought studies in medicine indicate that structured reasoning may support interpretability, but they also show the need to evaluate whether intermediate steps are clinically valid rather than merely fluent [15, 17, 27]. A rigorous evaluation should therefore score both the ranked differential and the reasoning pathway that produced it.

Real-time simulation

Real-time simulation should replay historical emergency department triage notes and vital signs in chronological order, withholding later labs, imaging, and clinician notes until after the model produces its initial differential. This design approximates the information constraints of triage and allows comparison between model output, retrospective outcomes, clinician differentials, admission decisions, and subsequent diagnostic revisions. Emergency LLM studies on triage, admission prediction, and documentation provide useful precedents for task-specific evaluation, but a differential diagnosis system should also measure latency, failure modes, retrieval relevance, and clinician usability under simulated ED time pressure [1-4, 21, 22]. The evaluation endpoint should be safe prioritization and decision support, not autonomous diagnosis.

Limitations

Technical limitations

The framework depends heavily on the completeness, currency, and local relevance of the retrieval knowledge base. If institutional protocols are missing, outdated, poorly indexed, or inconsistent with current practice, RAG may confidently ground the model in weak evidence rather than preventing hallucination. Chain-of-thought reasoning can also produce explanations that appear coherent while masking incorrect assumptions, a concern reflected across medical CoT and diagnostic LLM evaluations [15, 16, 31, 32]. Finally, real-time ED deployment requires balancing retrieval depth, reasoning detail, output verification, and latency, and this trade-off may vary across hospital infrastructure and patient acuity.

Clinical limitations

A triage-stage LLM cannot replace physical examination, longitudinal reassessment, bedside gestalt, laboratory testing, imaging, procedural judgment, or shared decision-making. The model may also be vulnerable to incomplete histories, biased documentation, atypical presentations, language barriers, and measurement error in vital signs. Studies of LLMs in complex diagnosis, emergency triage, and clinical reasoning indicate that these systems can support clinicians, but they remain insufficiently validated as autonomous diagnostic agents [5, 19, 20, 23]. Liability, clinician acceptance, workflow burden, and prospective safety evidence must therefore be addressed before such a system could be integrated into routine emergency care.

Conclusion

A large language model system for emergency department differential diagnosis should be designed around the realities of triage: limited time, incomplete data, physiologic uncertainty, and the need to identify dangerous conditions early. By combining triage notes, vital signs, retrieval-augmented generation, and structured reasoning, the proposed framework offers a pathway for generating ranked diagnostic hypotheses that remain clinically reviewable.

The key advantage of this approach is that it treats the LLM as a bounded reasoning component rather than an unconstrained diagnostic oracle. Retrieval provides grounding in curated clinical sources, chain-of-thought structure makes reasoning easier to inspect, and ranked output supports prioritization rather than premature closure.

The framework also has important limitations. It requires rigorous validation, strong governance, careful user-interface design, clinician override, and continuous monitoring for unsafe reasoning, biased outputs, and over-reliance.

Future work should focus on integration with emergency department information systems, prospective simulation, clinician-centered usability testing, and eventually controlled clinical trials. The goal is not to automate emergency diagnosis, but to create a safer, explainable, and real-time decision-support layer that helps clinicians reason under pressure.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Aug 2025 Revised: 18 Nov 2025 Accepted: 17 Jan 2026
Published online: 20 July 2026

References

Williams CY, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open. 2024;7(5):e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895.

Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res. 2024;26:e53297. https://doi.org/10.2196/53297.

Glicksberg BS, Timsina P, Patel D, Sawant A, Vaid A, Raut G, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. J Am Med Inform Assoc. 2024;31(9):1921-8.

Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, et al. Developing and evaluating large language model-generated emergency medicine handoff notes. JAMA Netw Open. 2024;7(12):e2448723. https://doi.org/10.1001/jamanetworkopen.2024.48723.

Yau JY, Saadat S, Hsu E, Murphy LS, Roh JS, Suchard J, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. J Med Internet Res. 2024;26:e60291. https://doi.org/10.2196/60291.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. https://doi.org/10.1038/s41586-023-06291-2.

Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80. https://doi.org/10.1001/jama.2023.8288.

Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI. 2024;1(1):AIp2300031. https://doi.org/10.1056/AIp2300031.

Shea YF, Lee CM, Ip WC, Luk DW, Wong SS. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. JAMA Netw Open. 2023;6(8):e2325000. https://doi.org/10.1001/jamanetworkopen.2023.25000.

Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8.

Wang C, Ong J, Wang C, Ong H, Cheng R, Ong D. Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. Ann Biomed Eng. 2024;52(5):1115-8. https://doi.org/10.1007/s10439-024-03447-4.

Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med. 2024;7(1):102. https://doi.org/10.1038/s41746-024-01047-2.

Ng KK, Matsuba I, Zhang PC. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. NEJM AI. 2025;2(1):AIra2400380. https://doi.org/10.1056/AIra2400380.

Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15.

Jeon S, Kim HG. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Comput Biol Med. 2025;196:110614. https://doi.org/10.1016/j.compbiomed.2025.110614.

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. https://doi.org/10.3390/medicina60010148.

Yang HS, Li J, Yi X, Wang F. Performance evaluation of large language models with chain-of-thought reasoning ability in clinical laboratory case interpretation. Clin Chem Lab Med. 2025;63(8):e199-201. https://doi.org/10.1515/cclm-2025-0211.

McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642(8067):451-7. https://doi.org/10.1038/s41586-025-08979-7.

Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.

Shah-Mohammadi F, Finkelstein J. Accuracy evaluation of GPT-assisted differential diagnosis in emergency department. Diagnostics (Basel). 2024;14(16):1779. https://doi.org/10.3390/diagnostics14161779.

Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. Am J Emerg Med. 2024;81:146-50. https://doi.org/10.1016/j.ajem.2024.05.021.

Haim GB, Saban M, Barash Y, Cirulnik D, Shaham A, Eisenman BZ, et al. Evaluating large language model assisted emergency triage: a comparison of acuity assessments by GPT-4 and medical experts. J Clin Nurs. 2024. https://doi.org/10.1111/jocn.17645.

Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr. 2024;178(3):313-5. https://doi.org/10.1001/jamapediatrics.2023.5757.

Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learn Health Syst. 2024;8(3):e10438. https://doi.org/10.1002/lrh2.10438.

Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. JAMA Netw Open. 2025;8(10):e2538427. https://doi.org/10.1001/jamanetworkopen.2025.38427.

Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. Digit Health. 2025;11:20552076251337177. https://doi.org/10.1177/20552076251337177.

Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Chain-of-thought reasoning improves ChatGPT's diagnostic accuracy in radiology. Can Assoc Radiol J. 2026;77(1):242-5. https://doi.org/10.1177/08465371251324567.

Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. Diagnostic performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023;308(1):e231040. https://doi.org/10.1148/radiol.231040.

Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health. 2024;2(1):4. https://doi.org/10.1186/s44247-024-00030-8.

Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of interleukin-1 receptor antagonist. J Transl Autoimmun. 2023;7:100213. https://doi.org/10.1016/j.jtauto.2023.100213.

Feldman MJ, Hoffer EP, Conley JJ, Chang J, Chung JA, Jernigan MC, et al. Dedicated AI expert system vs generative AI with large language model for clinical diagnoses. JAMA Netw Open. 2025;8(5):e2512994. https://doi.org/10.1001/jamanetworkopen.2025.12994.

Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open. 2026;9(4):e264003. https://doi.org/10.1001/jamanetworkopen.2026.4003.