MediLink: Safer Respiratory Triage Using Retrieval-Augmented Generation and Rule-Based Safety Gates

Jiayi Zheng[1,*], Surya Teja Nulu[2], Meryem Hassan Rafiq[1], Anand Ashika[1], and Aram Bahrini[1]
*jiayiz26@illinois.edu

[1]*Department of Business Administration*
Gies College of Business
University of Illinois at Urbana-Champaign
Champaign, IL, USA

[2]*Department of Computer Science*
Jarvis College of Computing and Digital Media
DePaul University
Chicago, IL, USA

***Abstract*—Large language models are increasingly used by patients seeking guidance for acute symptoms, particularly common respiratory complaints. However, unconstrained large language models can generate inaccurate or unsafe recommendations, particularly in triage settings where failure to recognize emergency symptoms may delay urgent care. This paper presents MediLink, a proof-of-concept respiratory triage framework that combines a deterministic Safety Gate with a retrieval-augmented generation pipeline grounded in guidance from the Centers for Disease Control and Prevention. The Safety Gate scans user input for red-flag symptoms and returns an immediate escalation message without invoking the language model, while the retrieval-augmented generation layer retrieves relevant Centers for Disease Control and Prevention guidance and injects it into the prompt for non-emergency cases. MediLink is evaluated on a synthetic dataset of 50 first-person patient cases across four conditions: Baseline, Retrieval-Augmented Generation Only, Safety Gate Only, and the full MediLink system. The evaluation examines emergency interception, false alarms, and response quality using human ratings and weighted Cohen's kappa. Results show that the Safety Gate consistently intercepts direct emergency descriptions, while indirect phrasing exposes the limitations of lexical rules; the retrieval-augmented generation module improves clarity while maintaining comparable actionability. These findings support the use of hybrid safety-aware architectures for constrained medical question answering in proof-of-concept settings while reinforcing the need for richer semantic safety mechanisms before real-world deployment.**

***Index Terms*—Retrieval-augmented generation, medical triage, rule-based safety, large language models, respiratory triage**

I. Introduction

Large language models (LLMs) have rapidly gained attention for their ability to generate fluent, context-sensitive responses across a wide range of applications [1]. In healthcare, they are increasingly being explored for tasks such as health communication, patient education, documentation, and clinical decision support [2], [3]. Yet the same properties that make them broadly useful—fluent language generation, broad domain coverage, and flexible reasoning—also create significant safety risks when they are used without safeguards in medical contexts. Recent clinical reviews emphasize that hallucinations, omissions, and unsupported claims remain major barriers to responsible deployment of LLMs in medicine [2], [4]. In triage-like scenarios, even a plausible but incorrect answer can be harmful if it fails to escalate a genuine emergency.

Respiratory symptom triage is a particularly relevant use case because patient descriptions are often brief, noisy, and expressed in first-person natural language rather than structured medical terminology. At the same time, public health guidance provides clear escalation thresholds for red-flag symptoms such as difficulty breathing, chest pain, or confusion [5], [6]. This creates an opportunity for a hybrid design that separates deterministic emergency handling from probabilistic response generation.

MediLink is designed as a proof-of-concept system for this setting. It combines a rule-based Safety Gate that performs pre-generation emergency detection with a retrieval-augmented generation (RAG) module that grounds non-emergency responses in Centers for Disease Control and Prevention (CDC) respiratory illness guidance. The goal is not to replace clinicians or validate autonomous diagnosis, but to demonstrate how a constrained architecture can improve safety and response quality in a controlled proof-of-concept setting.

The contributions of this paper are threefold. First, it presents a layered triage architecture that integrates deterministic safety logic with guideline-grounded generation. Second, it introduces a controlled evaluation protocol built around synthetic respiratory cases, including both direct and indirect emergency phrasing. Third, it analyzes how the Safety Gate and RAG layer contribute differently to safety and response quality in a proof-of-concept medical AI system.

II. Related Work

Research on LLMs in healthcare has consistently identified reliability and oversight as major barriers to deployment. Reviews of clinical LLMs highlight factual errors, hallucinations, transparency concerns, and the need for human oversight [2],

[3]. Recent assurance analyses in clinical decision-support settings further suggest that hallucination rates can remain substantial even with deterministic decoding, reinforcing the need for safeguards beyond prompting alone [4].

RAG is one of the most widely studied approaches for improving factual reliability in health-related LLM systems. By retrieving external evidence and injecting it into the prompt, RAG can better align responses with reference material, reduce unsupported claims, and improve source grounding. Systematic and narrative reviews in healthcare report growing interest in RAG for guideline interpretation, medical question answering, and patient education, while also noting the lack of standardized evaluation frameworks and the need for careful ethical deployment [7], [8]. Recent application studies also suggest that RAG can improve answer quality and source alignment in patient-facing medical chatbots [9].

In emergency and triage settings, the literature suggests that AI can support documentation, prioritization, and decision support, but current evidence remains heterogeneous and incomplete. A recent systematic review of AI triage systems in emergency departments found promise for efficiency and decision support, while also highlighting under-triage risk, variable accuracy, and the need for stronger validation and reporting standards [10]. These concerns are especially salient for patient-facing tools, where false reassurance can carry direct harm.

Despite progress in healthcare LLMs and RAG, fewer systems explicitly combine a deterministic safety layer with grounded generation in a simple, auditable triage pipeline. MediLink addresses this gap by placing a CDC-derived Safety Gate ahead of a CDC-grounded RAG module. This architecture is intentionally narrow in scope: it is designed as a proof-of-concept demonstration of modular safety and grounding rather than a clinically validated triage product.

III. System Design

A. System Overview

MediLink is a layered triage pipeline that processes user-submitted symptom descriptions through two sequential modules before generating a response. Upon receiving input, the system first evaluates the text against a rule-based Safety Gate. If no emergency indicators are detected, the input proceeds to the RAG module, which retrieves relevant CDC guideline chunks and injects them into a structured prompt for GPT–4. If the Safety Gate is triggered, the system returns a fixed emergency escalation message immediately, bypassing the LLM entirely. This ordering ensures that safety-critical cases are handled deterministically without relying on probabilistic model behavior. The overall system architecture is illustrated in Fig. 1.

B. Knowledge Base

The knowledge base consists of 16 document chunks extracted from five CDC respiratory illness guidance pages: About Respiratory Illnesses, Treatment of Respiratory Viruses, Precautions When Sick, Clinical Overview for Healthcare Providers, and People at Increased Risk. These chunks cover four clinically relevant content domains: symptom identification, high-risk population criteria, treatment timing recommendations, and emergency warning signs. CDC guidance was selected because it is authoritative, publicly available, and written in patient-accessible language that aligns with the system's user-facing design [5], [6].

Text chunks are embedded using the all-MiniLM-L6-v2 sentence-transformer model and indexed using FAISS with L2 distance. The resulting vector store contains 16 embeddings of 384 dimensions each.

C. RAG Module

When the Safety Gate does not trigger, the user input is encoded with the same embedding model and compared against the FAISS index to retrieve the top-3 most semantically similar CDC chunks ($k = 3$). These chunks are injected into a structured GPT–4o prompt as contextual reference material. The system prompt instructs the model to use only the provided CDC guidance, avoid unsupported information, and recommend consultation with a healthcare professional for individualized advice.

This design treats GPT–4o primarily as a grounded generation layer rather than a free-form medical advisor. Prior work in healthcare RAG motivates this choice: grounding external evidence can improve contextual relevance, support more transparent responses, and reduce unsupported content [7]–[9]. To maximize consistency across experimental conditions, generation is performed at a temperature of 0.0.

D. Safety Gate

The Safety Gate is a deterministic keyword-matching module that runs prior to any LLM call. It scans the user input for a predefined set of red-flag terms derived from CDC emergency warning signs, including expressions such as chest pain, difficulty breathing, coughing up blood, lips turning blue, confusion, and passing out [5], [6]. If any term is matched, the system immediately returns a fixed escalation message: "EMERGENCY: Please call 911 or go to the nearest emergency room immediately. Your symptoms may indicate a life-threatening condition."

Placing the Safety Gate before retrieval and generation guarantees that inputs containing predefined red-flag terms receive a consistent and unambiguous escalation response. This is valuable because, in safety-critical medical settings, prompt engineering alone is not a sufficient safeguard against hallucinations or unsafe under-escalation [2], [4]. However, the approach remains limited to lexical matching and can miss semantically equivalent emergency descriptions that do not contain the exact target phrases.

IV. Evaluation Methodology

A. Test Dataset

Evaluation was conducted on a synthetic dataset of 50 patient cases written in natural first-person language. Real patient data were not used because the objective of this work

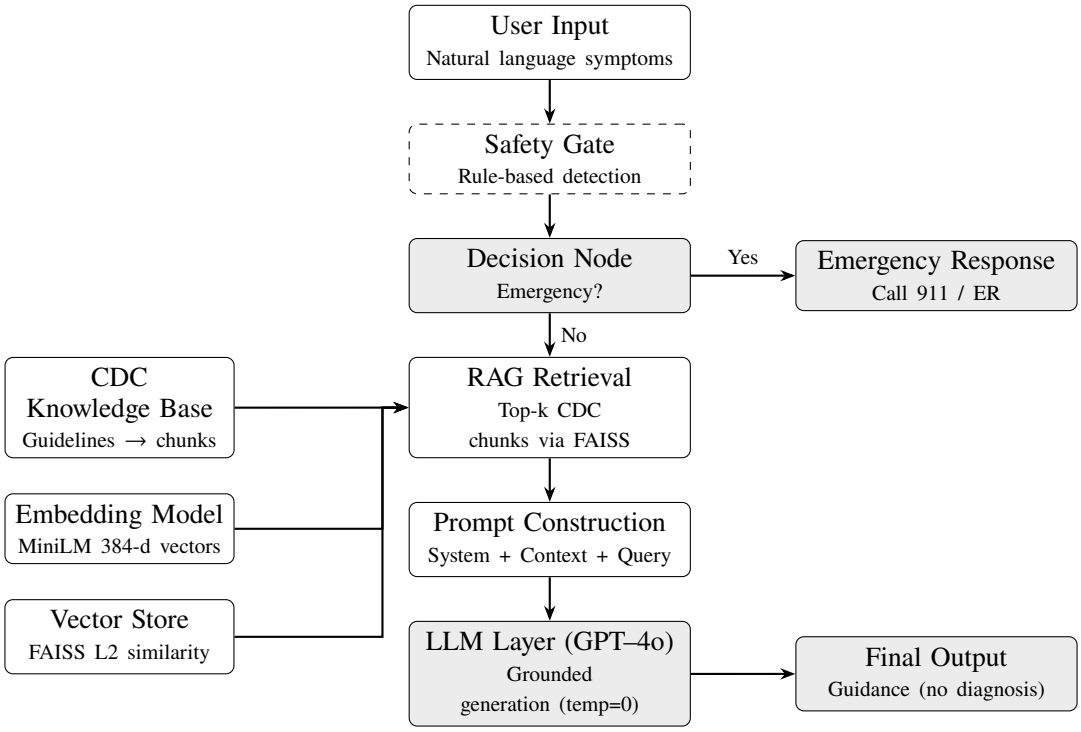

Fig. 1. MediLink system architecture. The pipeline combines a deterministic Safety Gate for emergency detection with a RAG module grounded in CDC guidance.

is to evaluate a proof-of-concept system rather than to conduct clinical validation. Synthetic data are therefore used to probe system behavior under controlled conditions while avoiding data-access constraints.

The dataset is divided into two subsets. Thirty non-emergency cases describe mild to moderate respiratory complaints such as cough, congestion, low-grade fever, and fatigue across both general and higher-risk populations. Twenty emergency cases describe symptoms requiring immediate attention. To test the boundary of the Safety Gate, emergency cases were deliberately written in two forms: eight direct cases that closely match the keyword list (e.g., "difficulty breathing" or "chest pain") and twelve indirect cases that describe the same urgency colloquially (e.g., "I feel like someone is sitting on my chest"). This split allows the evaluation to distinguish explicit lexical matches from semantically similar but lexically different inputs.

B. Experimental Conditions

All 50 cases were evaluated under four experimental conditions: (1) Baseline, where user input is passed directly to GPT–4o without the Safety Gate or retrieved context; (2) RAG Only, where the system retrieves the top-3 CDC chunks without applying the Safety Gate; (3) Safety Gate Only, where emergency detection is applied prior to generation but no CDC context is retrieved; and (4) MediLink (Full System), where the Safety Gate runs first and non-flagged inputs proceed through RAG-grounded generation.

C. Evaluation Metrics

Emergency performance is assessed using the Emergency Interception Rate and the False Alarm Rate. For all 20 emergency cases across all four conditions, two human raters independently assign an Urgency Score on a 1–5 ordinal scale: 1 for no danger mentioned, 2 for recommending a doctor without urgency, 3 for recommending care soon, 4 for explicit emergency-room guidance, and 5 for instructing the patient to call 911 or returning the fixed Safety Gate message. Scores of 4 or 5 are treated as successful interceptions.

For non-emergency cases, Baseline and MediLink responses are rated on two communication dimensions: clarity and actionability, both on a 1–3 scale. Clarity measures whether the response is easy to understand; actionability measures whether it provides concrete next steps. Because the communication-quality comparison focuses on end-to-end user-facing performance, these ratings are applied only to Baseline and MediLink responses. Weighted Cohen's kappa is used to assess inter-rater agreement because both scales are ordinal [11]. In addition, two to three representative case studies provide a side-by-side qualitative comparison of Baseline and MediLink outputs.

False Alarm Rate is defined as the proportion of non-emergency cases in which the system response contains explicit emergency escalation language, assessed using keyword-based detection over predefined escalation phrases.

V. Results and Discussion

A. Emergency Interception and False Alarms

Table I presents emergency interception rates and false alarm rates across all four conditions, broken down by overall performance and by direct versus indirect symptom phrasing.

TABLE I
Emergency Interception Rate and False Alarm Rate by Condition

Condition	Overall	Direct	Indirect	FAR
Baseline	55.0%	87.5%	33.3%	0.0%
RAG Only	95.0%	100.0%	91.7%	0.0%
Safety Gate Only	55.0%	100.0%	25.0%	3.3%
MediLink	90.0%	100.0%	83.3%	3.3%

For direct cases, both Safety Gate Only and RAG Only achieve perfect or near-perfect interception, while Baseline reaches 87.5%. This suggests that when symptoms are described using standard clinical terminology, even an unconstrained LLM can often recognize their urgency.

For indirect cases, the pattern diverges sharply. Safety Gate Only drops to 25.0%, confirming that rule-based keyword matching fails when patients use colloquial language. Baseline also performs poorly at 33.3%. By contrast, RAG Only achieves 91.7% on indirect cases, suggesting that grounding GPT–4o in CDC emergency criteria may help it recognize semantically equivalent but lexically different descriptions within this synthetic evaluation setting. MediLink achieves 83.3% on indirect cases, combining the deterministic reliability of the Safety Gate with RAG's broader semantic coverage. This suggests that semantic grounding may provide broader coverage than deterministic keyword matching when emergency descriptions are indirect or colloquial.

The marginally lower interception rate of MediLink compared to RAG Only on indirect cases (83.3% vs. 91.7%) indicates that the combined system does not outperform RAG Only on this specific metric. Instead, MediLink's advantage lies in combining semantic coverage with deterministic handling of inputs containing predefined red-flag terms. The Safety Gate's value should therefore not be interpreted primarily as a performance advantage in raw interception rate. Its contribution is determinism and consistency: for any input containing predefined red-flag terms, the Safety Gate guarantees an identical escalation response without invoking the LLM, independent of prompt variation, model version, or output stochasticity.

Inter-rater agreement for urgency scoring was strong across all conditions, with weighted Cohen's kappa ranging from $\kappa = 0.860$ to $\kappa = 1.000$, indicating that the scoring rubric was applied consistently by both raters.

The false alarm rate was low across all conditions. Baseline and RAG Only produced no false alarms. Safety Gate Only and MediLink each produced one false alarm (Case 17), in which a patient describing "coughing up clear mucus" triggered the Safety Gate due to partial substring matching with the "coughing up blood" keyword pattern. This illustrates a known limitation of exact string matching: semantically dissimilar inputs sharing lexical substrings can produce unintended escalations. In general, the low false alarm rate suggests that the Safety Gate does not substantially over-trigger on common non-emergency symptoms, although the observed error indicates that substring-based matching requires more careful keyword design.

B. Response Quality in Non-Emergency Cases

Table II presents average Clarity and Actionability scores for Baseline and MediLink across 30 non-emergency cases.

TABLE II
Response Quality Scores (Non-Emergency Cases)

Condition	Avg Clarity (1–3)	Avg Actionability (1–3)
Baseline	2.28	2.52
MediLink	2.59	2.48

MediLink responses showed a modest improvement in clarity over Baseline (2.59 vs. 2.28), suggesting that grounding responses in CDC guidelines produces more structured guidance. Actionability scores were comparable between conditions (2.48 for MediLink vs. 2.52 for Baseline), indicating that both systems provided similarly specific advice for non-emergency cases. This finding is consistent with the nature of the RAG module: CDC guidelines constrain response content and improve adherence to authoritative guidance, but do not necessarily increase the specificity of actionable steps beyond what a general-purpose LLM already produces for common respiratory complaints. In this setting, the main contribution of RAG appears to be better-structured and more guideline-aligned responses rather than more actionable recommendations.

Inter-rater agreement for quality scoring ranged from $\kappa = 0.613$ to $\kappa = 0.858$, with all values above 0.6, indicating moderate to strong agreement across the rated dimensions.

C. Representative Case Studies

Case 1 — Straightforward Non-Emergency (Case 7: Mild COVID, positive test). The Baseline response provided general self-care advice including rest, hydration, and over-the-counter medication options. The MediLink response specifically cited the CDC recommendation to stay home until symptoms improve and no fever for 24 hours without medication, and noted the 5-day precautionary period upon returning to normal activities. This illustrates how RAG grounds responses in specific CDC-sourced timelines rather than general advice.

Case 2 — High-Risk Patient (Case 21: 70-year-old with heart disease). The Baseline response gave standard cold and flu management advice without differentiating based on the patient's age or cardiac history. The MediLink response retrieved CDC high-risk population guidelines and emphasized the importance of seeking healthcare promptly, consistent with CDC guidance that early treatment is especially critical for older adults with underlying conditions.

Case 3 — Indirect Emergency (Case 43: "My thoughts are really foggy, I can barely hold a conversation"). This

case received a Baseline urgency score of 2 from both raters, indicating that the unconstrained LLM did not recognize the altered mental status as an emergency signal. The Safety Gate also failed to trigger, as the input contained no direct keyword matches. This case illustrates the core limitation of keyword-based approaches: semantically significant emergency language that does not match predefined patterns will go undetected.

VI. Limitations and Future Work

This study is subject to several limitations. First, the evaluation dataset is synthetically generated and does not capture the full diversity of real patient language, including idiomatic, culturally specific, abbreviated, or incomplete symptom descriptions. Conclusions should therefore be limited to the controlled conditions used in this proof-of-concept study.

Second, the Safety Gate's performance is directly bound by its keyword list. As demonstrated by the indirect emergency cases, genuine emergencies may go undetected when described using colloquial or semantically equivalent phrasing that does not lexically match predefined terms.

Third, the non-emergency evaluation emphasizes communication quality—specifically clarity and actionability—rather than clinical correctness. The research team does not claim that the system is clinically validated, diagnostically complete, or appropriate for autonomous decision-making. Results should not be interpreted as evidence of clinical efficacy or readiness for deployment.

Fourth, real-world validation would require substantially more diverse input data, independent clinician review, richer harm-oriented evaluation criteria, and testing across broader patient populations. Clinician review would also strengthen the evaluation of medical appropriateness, as the current rubric assesses communication quality rather than clinical correctness. Future work should incorporate semantic emergency detection, clinician-centered evaluation, and prospective assessment on real patient-reported data.

VII. Conclusion

This paper presented MediLink, a proof-of-concept AI-assisted triage system that integrates a deterministic Safety Gate with a RAG pipeline grounded in CDC respiratory illness guidelines. The system demonstrates how rule-based safety logic and knowledge-grounded language generation can be layered to improve reliability in patient-facing medical AI applications.

Experimental results show that the Safety Gate achieves perfect interception on cases whose phrasing directly matches predefined red-flag keywords, while the RAG module improves response clarity and shows stronger emergency recognition for indirect symptom descriptions in this synthetic test setting.

However, the Safety Gate's performance drops sharply for indirect phrasing (25.0%), highlighting that keyword-based detection alone is insufficient to provide robust coverage of unconstrained patient language. More broadly, the results suggest that deterministic safety logic and retrieval grounding contribute differently to system behavior: the former provides consistency for explicitly flagged inputs, while the latter appears to improve coverage of colloquial and indirectly expressed symptoms. All findings remain limited to synthetic evaluation conditions and do not establish real-world clinical performance.

Future work should explore semantic matching and learned classifiers to extend Safety Gate coverage, incorporate clinician evaluation for response quality assessment, and validate system behavior on real patient data. The implementation and evaluation data are publicly available on GitHub [12].

References

[1] A. Bahrini et al., "ChatGPT: Applications, Opportunities, and Threats," in 2023 Systems and Information Engineering Design Symposium (SIEDS), Apr. 2023, pp. 274–279.

[2] D. Roustan et al., "The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations," *JMIR Med Educ.*, vol. 11, e62847, 2025.

[3] S. Maity and M. J. Saikia, "Large Language Models in Healthcare and Medical Applications: A Review," *Healthcare*, vol. 13, no. 12, 2025.

[4] M. Omar et al., "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *npj Digital Medicine*, vol. 8, 2025.

[5] Centers for Disease Control and Prevention, "About Respiratory Illnesses," [Online]. Available: https://www.cdc.gov/respiratory-viruses/about/index.html. Accessed: Apr. 2026.

[6] Centers for Disease Control and Prevention, "Preventing Spread of Respiratory Viruses When You're Sick," [Online]. Available: https://www.cdc.gov/respiratory-viruses/prevention/precautions-when-sick.html. Accessed: Apr. 2026.

[7] L. M. Amugongo et al., "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, e0000877, 2025.

[8] O. K. Gargari et al., "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *Cureus*, vol. 17, no. 5, e84467, 2025.

[9] D. Baur, J. Ansorg, C.-E. Heyde, and A. Voelker, "Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study," *JMIR AI*, vol. 4, e80468, 2025.

[10] A. Z. Ahmed Abdalhalim et al., "Clinical Impact of Artificial Intelligence-Based Triage Systems in Emergency Departments: A Systematic Review," *Cureus*, vol. 17, no. 6, e86208, 2025.

[11] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical Therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[12] J. Zheng, S. Nulu, M. H. Rafiq, A., Ashika, and A. Bahrini, "Medilink-Paper," GitHub repository, 2026. [Online]. Available: https://github.com/jiayiz26/Medilink-Paper.