*Journal of the American Medical Informatics Association*, 2025, **32(6)**, 1032–1039
https://doi.org/10.1093/jamia/ocaf059
Advance access publication 12 April 2025
**Research and Applications**

Research and Applications

# Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation

**Siru Liu, PhD*[,1,2], Aileen P. Wright, MD, MS[1,3], Allison B. McCoy, PhD[1], Sean S. Huang, MD[1,3], Bryan Steitz, PhD[1], Adam Wright, PhD[1,3]**

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212, United States, [2]Department of Computer Science, Vanderbilt University, Nashville, TN 37240, United States, [3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States

*Corresponding author: Siru Liu, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue #1475, Nashville, TN 37212, United States (siru.liu@vumc.org)

## Abstract

**Objectives:** This study aims to develop and evaluate an approach using large language models (LLMs) and a knowledge graph to triage patient messages that need emergency care. The goal is to notify patients when their messages indicate an emergency, guiding them to seek immediate help rather than using the patient portal, to improve patient safety.

**Materials and Methods:** We selected 1020 messages sent to Vanderbilt University Medical Center providers between January 1, 2022 and March 7, 2023. We developed four models to triage these messages for emergencies: (1) Prompt-Only: the patient message was input with a prompt directly into the LLM; (2) Naïve Retrieval Augmented Generation (RAG): provided retrieved information as context to the LLM; (3) RAG from Knowledge Graph with Local Search: a knowledge graph was used to retrieve locally relevant information based on semantic similarities; (4) RAG from Knowledge Graph with Global Search: a knowledge graph was used to retrieve globally relevant information through hierarchical community detection. The knowledge base was a triage book covering 225 protocols.

**Results:** The RAG from Knowledge Graph model with global search outperformed other models, achieving an accuracy of 0.99, a sensitivity of 0.98, and a specificity of 0.99. It demonstrated significant improvements in triaging emergency messages compared to LLM without RAG and naïve RAG.

**Discussion:** The traditional LLM without any retrieval mechanism underperformed compared to models with RAG, which aligns with the expected benefits of augmenting LLMs with domain-specific knowledge sources. Our results suggest that providing external knowledge, especially in a structured manner and in community summaries, can improve LLM performance in triaging patient portal messages.

**Conclusion:** LLMs can effectively assist in triaging emergency patient messages after integrating with a knowledge graph about a nurse triage book. Future research should focus on expanding the knowledge graph and deploying the system to evaluate its impact on patient outcomes.

**Key words:** clinical decision support; large language model; message content; patient-doctor communication; primary health care; patient portal; knowledge graph; retrieval augmented generation.

## Introduction

Patients are increasingly using patient portals to communicate with their healthcare providers.[1] Between 2013 and 2018, the number of patient messages received by primary care providers (PCPs) increased by 110%.[2] This increase is partly due to increased out-of-pocket costs for in-person visits, leading patients to prefer online messages through patient portals.[3] The COVID-19 pandemic further intensified this trend, resulting in a 157% spike in patient messages—a trend that continues even post-pandemic.[4] At the same time, patients now appreciate direct and timely communication with their healthcare providers, with many anticipating near-instantaneous responses. However, response times often span several days.[5] While such delays may be manageable for non-emergencies, they can significantly endanger patient safety in emergency situations. An ongoing challenge is to efficiently identify emergencies, in which patients should be notified to call or seek emergency care rather than continue message exchanges. Currently, at Vanderbilt University Medical Center (VUMC), the Epic MyChart Patient Portal includes a disclaimer stating: "If you're having a medical emergency, call your provider, go to the nearest emergency department, or call 911." Despite this warning, patients occasionally send messages indicating potential medical emergencies to their healthcare providers. For example, a patient sent a message stating that they had a terrible headache with facial twitching and numbness (Figure 1). The patient portal is designed to streamline routine care and improve communications between doctors and patients, not to handle urgent situations. Consequently, there is a pressing need for an automated and efficient triage approach that can

**Received:** November 1, 2024; **Revised:** March 6, 2025; **Editorial Decision:** March 28, 2025; **Accepted:** April 1, 2025

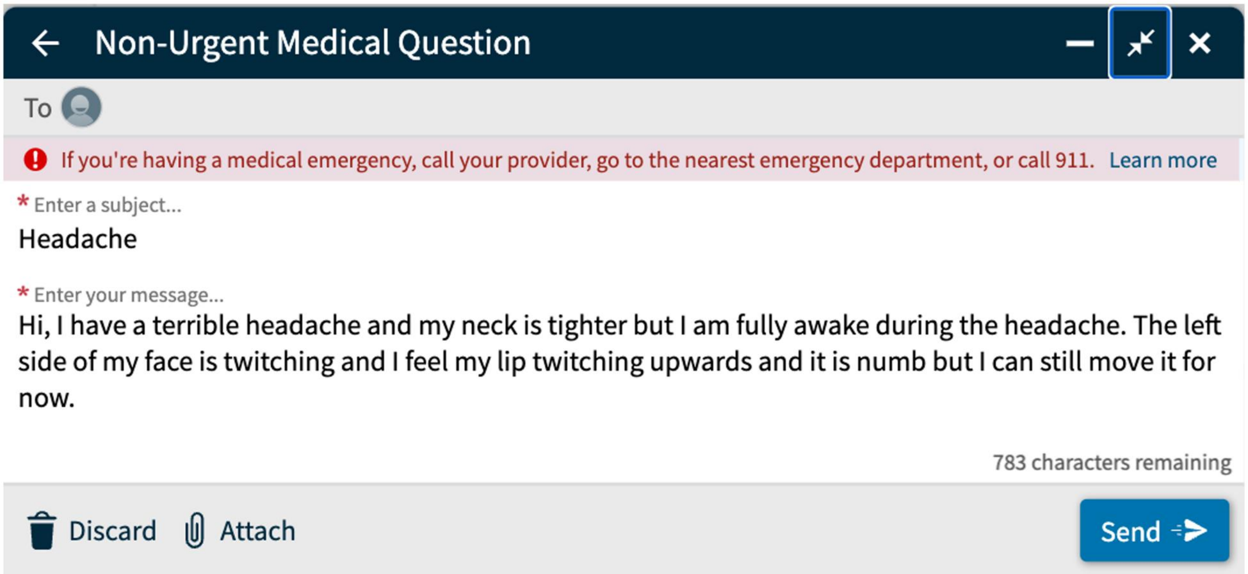


**Figure 1.** The screenshot of a patient message indicating a potential medical emergency. The form includes a warning to call 911 or visit the emergency department for emergencies, followed by text fields for "Enter a subject…" (filled in with "Headache") and "Enter your message…" (containing a detailed note about severe headache, neck tightness, left-side facial twitching, and lip numbness). The interface also shows buttons labeled "Discard," "Draft," "Attach," and "Send."

accurately identify patient messages requiring immediate medical attention. Such a system would help prevent patients from relying on the portal during the emergencies, ensuring patient safety.

LLMs are advanced AI models trained on vast amounts of text data to understand, reason, and generate human-like text across various domains.[6–8] LLMs show great potential for tasks such as medical question and answering,[7] clinical decision support analysis,[9,10] and assisting doctors and patients in writing messages.[11,12] However, challenges remain in applying LLMs to triage patient messages. A previous study evaluated GPT-3's performance in triaging 48 synthetic case vignettes and compared it to the performance of physicians and laypeople.[13] The study found that GPT-3's performance was inferior—even compared to laypeople—at correctly categorizing only 6 out of 12 emergency cases. This shortcoming could be attributed to the insufficient triage information for emergency cases in the training data for the LLM, or to an underperforming reasoning process regarding patient triage.[13] It is also reflective of the clinical complexity of patient triage. Moreover, LLMs sometimes generate fabricated or inaccurate information, a phenomenon known as "hallucination," which is particularly problematic in clinical settings.[14–16]

Integrating domain-specific knowledge into LLMs is a promising approach to address these challenges. In healthcare, knowledge graphs can construct medical concepts and relationships that improve clinical reasoning and decision-making.[17] Combining knowledge graphs with LLMs can improve inference capabilities.[18,19] Providing comprehensive, high-quality context can help mitigate hallucination.[20,21] Moreover, retrieval augmented generation (RAG) has been shown to improve accurate and more useful results to novel queries.[22–24] Integrating external knowledge, for example, a nurse triage book with over 200 protocols for common concerns and symptoms, might further support the emergency triage task.[25] RAG can be categorized into naïve and advanced approaches. Naïve RAG employs similarity metrics to extract relevant text snippets from documents, which are then directly incorporated into the LLM's prompt as contextual information.[26] Naïve RAG approaches often suffer from the "lost-in-the-middle" phenomenon, where important information is missed or overlooked in large and unstructured document sets.[27,28] Integrating knowledge graphs might improve the quality of retrieved information, leading to more accurate and reliable outputs, which is crucial for clinical tasks such as triage.[29,30] The objective of this research is to develop and evaluate an evidence-based LLM-powered triage system for emergency patient messages by integrating knowledge graphs to enhance accuracy and reliability.

## Materials and methods

This study took place at VUMC and was exempted from review by the Vanderbilt University Institutional Review Board. An overview of the study design is presented in Figure 2. We collected a dataset consisting of messages sent by patients to their PCPs and the corresponding replies via the My Health at Vanderbilt patient portal (Epic's MyChart, Verona, WI) between January 1, 2022 and March 7, 2023. We also collected the subject of each patient message. Protected health information (PHI) was removed from the messages using an automatic de-identification tool to ensure patient privacy,[31] but the content was otherwise left unchanged. We manually reviewed physician responses and randomly selected a set of non-emergency messages. Regarding emergency messages, we observed that healthcare providers were less likely to reply to patients instructing them to go to the emergency room through patient portal; instead, they often called patients to inform them of the emergency and documented it in a "Telephone Encounter" note in the Epic electronic health record (EHR). Other types of notes might also include information to identify emergencies in patient messages. In this study, to select messages indicating potential emergencies, we extracted notes written within 24 h after a patient message,

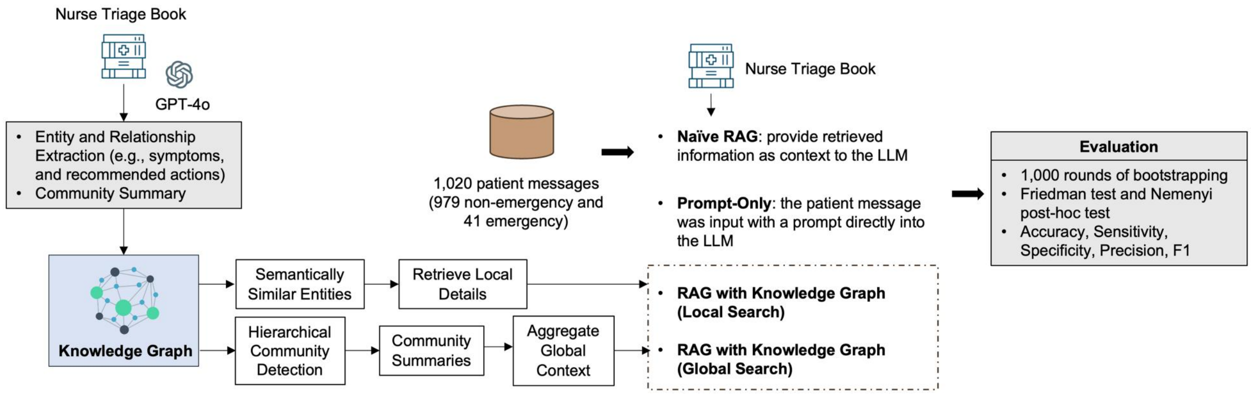


**Figure 2.** Study overview. A flowchart showing a study overview where a knowledge graph, developed using a GPT-4–powered nurse triage book, feeds into three RAG methods (naive, local, global) and a prompt-only approach, with all outputs evaluated via statistical tests on key performance metrics. (RAG: retrieval augmented generation, LLM: large language model).

with a note type marked as "Telephone Encounter." Within these notes, we searched for the term "go to ER" and other synonyms to identify potential emergency cases. An internal medicine physician then reviewed the dataset and confirmed 41 emergency patient messages. The total dataset included 1020 messages (979 non-emergency and 41 emergency).

The LLM used was GPT-4o, deployed via Microsoft Azure in a secure environment at VUMC to ensure patient data protection. The setting parameters are listed in the Appendix S1. Although this version of GPT-4o is approved for use with PHI and does not store prompts for future model training, we de-identified the messages before analyzing them for an additional layer of privacy protection. For each patient message, we applied four methods to generate responses: (1) Prompt-Only: The patient message was input with a prompt directly into the LLM without additional context; (2) Naïve RAG: Relevant information was retrieved using semantic similarities and provided as context to the LLM; (3) RAG from Knowledge Graph with Local Search: A knowledge graph was used to retrieve locally relevant information based on semantic similarities; (4) RAG from Knowledge Graph with Global Search: A knowledge graph was used to retrieve globally relevant information through hierarchical community detection. In particular, the local search was performed by identifying semantically similar entities to the patient message within the knowledge graph. Associated text chunks were retrieved, assessed for relevance by the LLM, and condensed into a single context window. On the other hand, for the global search, we first partitioned the knowledge graph into hierarchical communities using the Leiden Algorithm.[32] Community detection is critical for understanding large and complex graphs, and the Leiden algorithm is a state-of-the-art approach that guarantees connected communities while converging to partitions where all subsets are locally optimally assigned. Additionally, its fast local move approach makes it computationally more efficient, providing improved partition quality. Summaries were generated for communities at each level. The LLM evaluated these summaries for their relevance in assessing the patient messages. Relevant community summaries were combined until reaching the maximum context length allowed by the LLM. We implemented the RAG from Knowledge Graph approach using Microsoft's GraphRAG package in Python.[33] The prompts are listed in Appendix S2.

**Table 1.** Categorization of patient messages.

| Number of patient messages | |
|---|---|
| Medication and prescription management | 235 |
| Lab and test results | 215 |
| Chronic condition and pain management | 207 |
| General follow-up and appointments | 158 |
| Infections and immunizations | 93 |
| Blood pressure and cardiovascular concerns | 82 |
| Endocrinology and metabolic health | 57 |
| Mental health and psychiatric support | 50 |
| Other/miscellaneous | 80 |

We used GPT-4o to develop a knowledge graph based on the nursing telephone triage book, published in 2021.[25] The book includes adult, pediatric, geriatric, and maternal/child concerns, covering 225 protocols for common symptoms, disorders, and medical emergencies. Each protocol has been peer-reviewed by clinical experts to ensure accuracy and up-to-date content. The protocols are primarily symptom-based (eg, anxiety, back pain) but also include disease-based entries (eg, COVID-19 exposure, avian influenza). To construct the knowledge graph, we extracted entities such as symptoms, and recommended actions from the protocols. Relationships between these entities were established based on the guidelines provided in the triage book. Nodes in the graph represent medical concepts, while edges represent the relationships or protocols connecting them. To conduct a preliminary evaluation of the knowledge graph, we randomly selected 100 edges and compared them to the original text to assess their accuracy.

For the naïve RAG approach, we used the Chroma vector database[34] to store embeddings of the knowledge base. Text from the nursing triage guide was segmented into chunks of up to 2000 tokens. When processing a patient message, we computed its embedding and retrieved the top five most similar chunks based on semantic similarities. These retrieved chunks were included as additional context in the prompt provided to the LLM.

We conducted statistical analyses to compare the effectiveness of the four methods. We performed 1000 rounds of bootstrapping to estimate the sampling distributions of the performance metrics. The Friedman test, a non-parametric test for comparing more than two related groups, was used to determine if there were significant differences among the

**Table 2.** Model performance.

| Model | Accuracy | Sensitivity | Specificity | Precision | F1 |
|---|---|---|---|---|---|
| Prompt-only | 0.91 [0.89, 0.93] | 0.73 [0.57, 0.86] | 0.92 [0.90, 0.93] | 0.27 [0.19, 0.35] | 0.39 [0.29, 0.49] |
| Naïve RAG | 0.95 [0.94, 0.96] | 0.83 [0.71, 0.93] | 0.96 [0.95, 0.97] | 0.45 [0.34, 0.57] | 0.59 [0.47, 0.68] |
| RAG from knowledge graph (local search) | 0.98 [0.97, 0.98] | 0.95 [0.88, 1.00] | 0.98 [0.97, 0.98] | 0.63 [0.51, 0.74] | 0.76 [0.67, 0.84] |
| RAG from knowledge graph (global search) | 0.99 [0.98, 1.00] | 0.98 [0.92, 1.00] | 0.99 [0.98, 1.00] | 0.80 [0.68, 0.91] | 0.88 [0.80, 0.95] |

methods.[35,36] If the Friedman test indicated significant differences, we conducted a Nemenyi post-hoc test to identify which pairs of methods differed significantly.[37] Statistical significance was set at $P < .005$.

## Results

The average length of patient messages was 427 characters, sent by 954 unique patients. Based on their subjects, the messages were categorized into nine types, reflecting various healthcare needs (Table 1). The most common category was Medication and Prescription Management, with 235 messages. These messages frequently included requests for prescription refills, new medications, and inquiries about specific therapies. Many patients sought dosage adjustments or changes in their medication routines. The second most common category, Lab and Test Results, accounted for 215 messages. Patients primarily discussed lab results, including about comprehensive metabolic panels, lipid panels, A1C tests, and hormone tests, as well as imaging diagnostics such as CT scans, X-rays, and MRIs. Some messages involved follow-up questions about cancer screenings and other specialized tests.

The knowledge graph constructed from the nursing triage protocols contained 29 680 entities, 5142 relationships, and 249 communities. The preliminary evaluation of its quality, based on a random sample of 100 edges, yielded an accuracy score of 0.95. For example, one error was identified for the edge from "PELVIC INFLAMMATORY DISEASE" to "URINARY TRACT INFECTION." The associated description was, "UTIs can sometimes lead to PID if the infection spreads," yet this relationship was not explicitly stated in the original text. Notably, this edge had the lowest weight (7) among all the sampled edges. Table 2 summarizes the performance metrics of four models: Prompt-Only, Naïve RAG, RAG from Knowledge Graph (Local Search), and RAG from Knowledge Graph (Global Search). The prompt-only model, while achieving an accuracy of 0.91 [0.89, 0.93], showed limitations in sensitivity (0.73 [0.5, 0.86]) and precision (0.27 [0.19, 0.35]), indicating a lower capability to correctly identify emergency messages and a higher rate of false positives. In comparison, the naïve RAG model demonstrated moderate improvements across all metrics, particularly in sensitivity (0.83 [0.71, 0.93]) and precision (0.45 [0.34, 0.57]). The RAG from Knowledge Graph with local search model further enhanced performance, reaching 0.98 [0.97, 0.98] in both accuracy and specificity, along with a precision of 0.63 [0.51, 0.74]. The RAG from Knowledge Graph with global search achieved the highest performance across all metrics. This model also showed the highest precision (0.80 [0.68, 0.91]) and F1 score (0.88 [0.80, 0.95]), indicating its effectiveness in reliably detecting emergency cases while minimizing false positives. The Friedman tests revealed significant differences among all models for each metric ($P = .001$). In the subsequent Nemenyi post-hoc tests, each model was found to be significantly different from the others ($P = .001$).

The RAG from Knowledge Graph (global search) model misclassified 10 non-emergency messages (0.98%) as emergencies and 1 emergency message (0.1%) as a non-emergency. Table 3 provides examples of misclassified patient messages, and the responses generated by each model. For instance, an emergency message describing symptoms of a possible drug interaction with the COVID-19 medication was classified as a non-emergency by the RAG from Knowledge Graph with global search. This misclassification likely occurred due to the absence of specific COVID-19 drug-drug interaction resources in the knowledge graph. In contrast, the naïve RAG model correctly identified the case as an emergency by matching symptoms such as blurry vision, which indicated an urgent concern. Another example involved a non-emergency patient message, which requested a colonoscopy due to observed blood in the stool and a family history of colon cancer, and was correctly classified as a non-emergency only by the RAG from Knowledge Graph with global search. Although the patient expresses concern about blood in the stool and a family history of colon cancer, the absence of acute symptoms (eg, severe pain, dizziness, or vomiting blood) aligns with nurse triage protocols that classify such cases as non-emergencies.[25] The other three models, prompt only, naïve RAG, and RAG from Knowledge Graph with local search, focused on the blood and family history of cancer, leading them to incorrectly classify it as an emergency. The final example was about the low iron level. The patient directly emphasized urgency in the message with the phrase "It is urgent." The prompt-only model misclassified it as an emergency message, while other three models successfully filtered it out. In addition, in a random sample of 30 patient messages, we did not identify any instances where a correct judgment was supported by a flawed reasoning process.

## Discussion

The results of our study indicated that incorporating a knowledge graph-based RAG approach significantly enhances the performance of LLMs in triaging patient messages which indicate a potential emergency. Among all models tested, the RAG from Knowledge Graph with global search yielded the highest accuracy, sensitivity, and specificity, demonstrating its efficacy in accurately identifying emergency situations in patient messages.

The traditional LLM without any retrieval mechanism underperformed compared to models with RAG, which aligns with the expected benefits of augmenting LLMs with

**Table 3.** Examples of misclassified patient messages. The text shown in bold indicates the key differences between methods.

Example 1

Emergency Patient Message:

I went to the walk-in clinic for COVID and got Paxlovid. When I took it, I started feeling really anxious and restless, had muscle spasms and had blurry vision with small pupils. I read online that buspirone and Paxlovid can interact with each other. What should I do?

Prompt-Only:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

**Could you please confirm the dosage and frequency of both Paxlovid and Buspar that you are currently taking?**

Naïve RAG:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. **Based on the symptoms you described, including anxiousness, muscle spasms, restlessness, and small pupils with blurry vision, it is important to seek emergency care immediately.** Please call 911 or go to the nearest emergency room right away.

RAG from Knowledge Graph (Local Search):

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

**First, can you please confirm if you are currently experiencing any severe or life-threatening symptoms such as difficulty breathing, chest pain, or severe allergic reactions?**

RAG from Knowledge Graph (Global Search):

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

To help your physician understand your situation better, could you please provide the following information: **How long have you been experiencing these symptoms (anxiousness, muscle spasms, restlessness, and small pupils with blurry vision)?**

Example 2:

Non-Emergency Patient Message:

Over the last two days I have noticed blood in my stool, and I have a family history of colon cancer (my dad passed away from it). It might be a hemorrhoid, but I would feel better if I got a colonoscopy, since it has been several years since I had mine. Could you please order a colonoscopy for me as soon as possible?

Prompt-Only:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. **Blood in the stool can be a serious concern, especially with a family history of colon cancer. I recommend you seek immediate medical attention.** Please call 911 or go to the nearest emergency room right away. Your safety is the most important thing.

Naïve RAG:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

**Given that you have noticed blood in your stool and have a family history of colon cancer, I recommend you seek emergency care immediately.** Please call 911 or go to the nearest emergency room.

RAG from Knowledge Graph (Local Search):

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

**Given the presence of blood in your stool and your family history of colon cancer, it is important to seek immediate medical attention**. Please call 911 or go to the nearest emergency room right away.

RAG from Knowledge Graph (Global Search):

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

I understand that you have noticed blood in your stool for the past two days and have a family history of colon cancer. You mentioned that it has been several years since your last colonoscopy. To help your physician better understand your situation, **could you please provide more details about the amount of blood you have noticed and if you have experienced any other symptoms such as abdominal pain or weight loss?**

Example 3:

Non-Emergency Patient Message:

It is urgent, would you please call me at [PHONE NUMBER]. My iron is very low. I feel no energy. I want to sleep all day and night, and like eating ice and carrots. I must check my iron now. I don't know what to do or who to see. Thank you.

Prompt-Only:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. **Based on what you've shared, you might require immediate medical attention**. I recommend that you call 911 or go to the nearest emergency room right away for your symptoms.

Naïve RAG:

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

Your symptoms of extreme exhaustion, constant need to sleep, and craving non-food items like ice and carrots can be concerning. Let's proceed with gathering more information to help your physician understand your situation better. **Can you please confirm if you have any other symptoms or medical conditions that might be relevant?**

RAG from Knowledge Graph (Local Search):

This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.

Based on your message, it sounds like you are experiencing extreme exhaustion, constant need for sleep, lack of energy, and cravings for ice and carrots, which you believe are due to low iron levels.

Before we proceed, I need to check if you are experiencing any conditions that might require immediate medical attention. **Are you currently feeling confused, too weak to stand, or experiencing any signs of a stroke such as sudden-onset numbness, difficulty walking, talking, swallowing, or thinking?**

(continued)

Table 3. (continued)

RAG from Knowledge Graph (Global Search):
This is PAIGE. I'm an AI helper that will assist you in sending a message to your physician. I will ask you some follow-up questions to clarify important information and help ensure your physician will be able to accurately answer your question the first time.
Based on your message, it sounds like you are experiencing extreme exhaustion, constant need for sleep, lack of energy, and cravings for ice and carrots, which you believe are due to very low iron levels. Before we proceed, I need to check if you are experiencing any conditions that might require immediate medical attention. **Are you currently:**
**- Confused or too weak to stand?**
**- Experiencing signs of meningitis such as fever, confusion, headache, vomiting, stiff neck, or a red or blood-colored flat rash?**
**- Experiencing signs of a stroke such as sudden-onset numbness or tingling, difficulty walking, talking, swallowing, or thinking?**
Please let me know if any of these apply to you. If not, we can continue preparing your message to your physician.



domain-specific knowledge sources. The naïve RAG model showed improvements over the prompt-only method but was less effective than the RAG from Knowledge Graph approaches, particularly in emergency message classification. The ability of RAG from Knowledge Graph to identify semantically related entities and summarize them efficiently led to more accurate and context-aware responses, mitigating the "lost-in-the-middle" phenomenon described in previous studies.[27] Incorporating a knowledge graph as part of the RAG approach provided a structured way to represent relationships between medical concepts, which in turn enhanced the LLM's inference capabilities. By systematically organizing entities and relationships into communities, the global search model was able to provide a more comprehensive understanding of the patient's condition, leading to improved triage performance.

The low performance of the prompt-only model matched previously reported outcomes,[38] as did the improvement of the naïve RAG model over the prompt-only model.[39] These studies found a substantial reduction in under-triage rates from 39% to 8% when using GPT-4 without vs with RAG. While these studies used a small number of vignettes or simulated scenarios (eg, <150),[40] we were able to test on a large set of real patient messages with validated labels based on physicians' actual responses and with verification by an additional internal medicine physician. Another strength was our development of a knowledge graph using a standard telephone triage book as the knowledge base for the RAG. Our results suggest that providing this external knowledge, especially in a structured manner and in community summaries, can improve LLM performance in triaging patient portal messages.

While we integrated a book for nursing triage, a simple and more protocolized medical activity that lends itself well to a RAG, future research could expand the knowledge base to include clinical guidelines, local triage policies, and external APIs, such as RxNorm for medication information. This expansion could further enhance the RAG from Knowledge Graph model's ability to handle a broader range of medical scenarios and improve its generalizability. Another approach could be developing a multi-agent system. Instead of relying only on pre-stored knowledge, this system could assign agents various tasks, allowing them to work collaboratively to handle patient messages more comprehensively. For instance, agents could perform real-time searches on the latest clinical guidelines and extract patient EHR information through FHIR, providing accurate and personalized responses.

Finally, the implementation and evaluation of such tools in the patient portal is an important direction for future work.[41] Currently, when a patient portal message is received by the clinic and determined to indicate a potential emergency, this message must be responded to by telephone, since messages are not always read promptly by the patient. This process can be time-consuming and lead to delays in the patient receiving appropriate care as the clinic tries to reach the patient. An AI tool which automatically detected a potential emergency in the text of a patient portal message could display an immediate warning to the patient at the time of sending the message, and suggest contacting the clinic by telephone or seeking emergency medical care for the condition they are describing. Even in the case of non-emergency messages, an LLM-powered triage approach could place messages into categories based on urgency, such as needing to be seen within one day, within one week, at the next available appointment, or appropriate for self-care. Such a system could provide this feedback directly to patients[13] and be integrated into EHR systems to flag messages which should be reviewed urgently by the healthcare provider. These systems could help improve effective management of inbox messages, which is a pressing issue for PCPs who process approximately 150 messages per day.[42,43]

### Limitations

One limitation of our study is the reliance on a nursing telephone triage book as a single source for knowledge graph development, which may not cover all possible emergency scenarios. Expanding the knowledge graph with additional sources, such as clinical guidelines and expert knowledge, could further enhance the model's performance. Additionally, the evaluation was performed using a historical dataset, and real-world testing is needed to validate the system's effectiveness in clinical practice.[44] Moreover, LLMs may occasionally produce correct answers while relying on flawed reasoning processes.[44] To effectively implement this tool in a clinical workflow, further detailed evaluations focusing on the reasoning process are warranted. A potential direction is to identify hidden flaws and penalize them in the reinforcement learning process to fine-tune a specialized LLM for emergency information detection.

## Conclusion

By using RAG from a knowledge graph, this study demonstrates that LLMs can effectively assist in detecting emergencies in patient messages, offering significant improvements over LLMs without RAG and naïve RAG approaches. Future research should focus on expanding the knowledge graph and deploying the system in real clinical environments to evaluate its impact on patient outcomes.

## Author contributions

Siru Liu (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Aileen P. Wright (Conceptualization, Data curation, Methodology, Resources, Writing—review & editing), Allison B. McCoy (Conceptualization, Methodology, Resources, Supervision, Writing—review & editing), Sean S. Huang (Data curation, Resources, Validation, Writing—review & editing), Bryan D. Steitz (Methodology, Writing—review & editing), and Adam Wright (Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing—review & editing)

## Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## Funding

This work was supported by NIH [R00LM014097-02 and R01LM013995-01].

## Conflicts of interest

None declared.

## Data availability

The prompts were reported in the Appendix.

## References

1. Huang M, Fan J, Prigge J, et al. Characterizing patient-clinician communication in secure medical messages: retrospective study. *J Med Internet Res*. 2022;24:e17273. https://doi.org/10.2196/17273
2. North F, Luhman KE, Mallmann EA, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? *JMIR Med Inform*. 2020;8:e16521. https://doi.org/10.2196/16521
3. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. *J Gen Intern Med*. 2022;37:4002-4003. https://doi.org/10.1007/s11606-022-07766-0
4. Holmgren AJ, Downing NL, Tang M, et al. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *J Am Med Inf Assoc*. 2022;29:453-460. https://doi.org/10.1093/jamia/ocab268
5. Shimada SL, Petrakis BA, Rothendler JA, et al. An analysis of patient-provider secure messaging at two Veterans Health Administration medical centers: message content and resolution through secure messaging. *J Am Med Inf Assoc*. 2017;24:942-949. https://doi.org/10.1093/jamia/ocx021
6. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901. https://doi.org/10.48550/arXiv.2005.14165
7. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589-596. https://doi.org/10.1001/jamainternmed.2023.1838
8. Alapati R, Campbell D, Molin N, et al. Evaluating insomnia queries from an artificial intelligence chatbot for patient education. *J Clin Sleep Med*. 2024;20:583-594. https://doi.org/10.5664/jcsm.10948
9. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inf Assoc*. 2023;30:1237-1245. https://doi.org/10.1093/jamia/ocad072
10. Liu S, McCoy AB, Wright AP, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. *J Am Med Inf Assoc*. 2024;31:1388-1396. https://doi.org/10.1093/jamia/ocae041
11. Liu S, Wright AP, Mccoy AB, et al. Using large language model to guide patients to create efficient and comprehensive clinical care message. *J Am Med Inf Assoc*. 2024;31:1665-1670. https://doi.org/10.1093/JAMIA/OCAE142
12. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. *J Am Med Inf Assoc*. 2024;31:1367-1379. https://doi.org/10.1093/jamia/ocae052
13. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digit Health*. 2024;6:e555-61.e561. https://doi.org/10.1016/S2589-7500(24)00097-9
14. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2401.11817
15. Zhang Y, Li Y, Cui L, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. 2023, preprint: not peer reviewed. https://https://arxiv.org/abs/2309.01219
16. Huang L, Yu W, Ma W, et al. a survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2311.05232
17. Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records. *Sci Rep*. 2017;7:5994. https://doi.org/10.1038/s41598-017-05778-z
18. Lo P-C, Tsai Y-H, Lim E-P, et al. On exploring the reasoning capability of large language models with knowledge graphs. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2312.00353
19. Tang W, Qiao X, Zhao X, et al. HW-TSC at TextGraphs-17 shared task: enhancing inference capabilities of LLMs with knowledge graphs. In: *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*. Bangkok, Thailand: Association for Computational Linguistics; 2024:131-136.
20. Li X, Zhao R, Chia YK, et al. Chain-of-knowledge: grounding large language models via dynamic knowledge adapting over heterogeneous sources. In: *12th International Conference on Learning Representations*. Vienna, Austria: International Conference on Learning Representations, ICLR; 2024.
21. Peng B, Galley M, He P, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2302.12813
22. Xiong G, Jin Q, Wang X, et al. Improving retrieval-augmented generation in medicine with iterative follow-up questions. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2408.00727
23. Huang H, Wang S, Liu H, et al. Benchmarking large language models on communicative medical coaching: a dataset and a novel system. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics; 2024:1624-1637. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.94
24. 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:ocaf008. https://doi.org/10.1093/JAMIA/OCAF008
25. Briggs JK. *Telephone Triage Protocols for Nurses*. Wolters Kluwer Health; 2021.

26. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. 2023, preprint: not peer reviewed. https://arxiv.org/abs/2312.10997
27. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *Trans Assoc Comput Linguist*. 2024;12:157-173. https://doi.org/10.1162/tacl_a_00638
28. Xu P, Ping W, Wu X, et al. Retrieval meets long context large language models. In: *12th International Conference on Learning Representations*. Vienna, Austria: International Conference on Learning Representations, ICLR; 2024.
29. Pan S, Luo L, Wang Y, et al. Unifying large language models and knowledge graphs: a roadmap. *IEEE Trans Knowl Data Eng*. 2024;36:3580-3599. https://doi.org/10.1109/TKDE.2024.3352100
30. Sansford H, Richardson N, Maretic HP, et al. GraphEval: a knowledge-graph based LLM hallucination evaluation framework. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2407.10793
31. Chambon PJ, Wu C, Steinkamp JM, et al. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *J Am Med Inf Assoc*. 2023;30:318-328. https://doi.org/10.1093/jamia/ocac219
32. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9:5233. https://doi.org/10.1038/s41598-019-41695-z
33. Edge D, Trinh H, Cheng N, et al. From local to global: a graph rag approach to query-focused summarization. 2024, preprint: not peer reviewed. https://arxiv.org/abs/2404.16130
34. Chroma. Accessed March 2, 2025. https://www.trychroma.com/
35. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024;14:6086. https://doi.org/10.1038/s41598-024-56706-x
36. Demsar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1-30.
37. Nemenyi PB. *Distribution-Free Multiple Comparisons*. Princeton University; 1963.
38. Franc JM, Cheng L, Hart A, et al. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *Can J Emerg Med*. 2024;26:40-46. https://doi.org/10.1007/s43678-023-00616-w
39. Yazaki M, Maki S, Furuya T, et al. Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model. *Prehospital Emerg Care*. 2025;29:203-209. https://doi.org/10.1080/10903127.2024.2374400
40. Masanneck L, Schmidt L, Seifert A, 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
41. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res*. 2023;25:e48568. https://doi.org/10.2196/48568
42. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. *J Am Med Inf Assoc*. 2021;28:923-930. https://doi.org/10.1093/jamia/ocaa229
43. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med*. 2017;15:419-426. https://doi.org/10.1370/afm.2121
44. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *NPJ Digit Med*. 2024;7:190. https://doi.org/10.1038/s41746-024-01185-7





Journal of the American Medical Informatics Association, 2025, 32, 1032–1039
https://doi.org/10.1093/jamia/ocaf059
Research and Applications