*applied sciences*

MDPI

*Article*

# TRAG: A Rule Retrieval-Augmented Generation Framework for Emergency Triage

**Luluah Alhusain**

Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia; lalhusain@ksu.edu.sa

**Abstract**

Accurate emergency triage is critical for patient safety and efficient resource allocation. Large language models (LLMs) have shown promise in clinical reasoning tasks; however, their predictions may be inconsistent when not grounded in structured clinical knowledge. This study proposes TRAG (Triage Retrieval-Augmented Generation), a domain-specific framework that integrates rule-based knowledge retrieval with LLM reasoning to support Emergency Severity Index (ESI) prediction. TRAG retrieves triage rules encoded as if–then logic from an ESI knowledge base and incorporates them into the model prompt to guide ESI prediction. The framework evaluated multiple LLMs under different retrieval settings (Top-5, Top-10, Top-15) and compared with zero-shot baselines on 100 curated triage cases. Performance was assessed using accuracy, precision, recall, F1-score, and Quadratic Weighted Kappa (QWK). Results show that retrieval-augmented prompting improves classification performance, particularly for lower-performing models. For example, GPT-3.5 accuracy increased from 0.45 to 0.68 and QWK from 0.67 to 0.82 under the Top-15 setting. Improvements were also observed in reducing under-triage in several configurations, while higher-performing models demonstrated more modest and configuration-dependent gains. These findings suggest that integrating structured clinical rules within a retrieval-augmented framework can enhance the consistency and reliability of LLM-based triage prediction. The proposed TRAG framework highlights the potential of combining structured clinical knowledge with generative models to support safer and more interpretable decision-making in emergency care.

**Keywords:** emergency triage; large language models; retrieval-augmented generation; clinical decision support; emergency severity index; artificial intelligence

Check for updates

Academic Editors: Giacomo Cappon and Luca Cossu

Received: 12 February 2026
Revised: 24 April 2026
Accepted: 28 April 2026
Published: 5 May 2026

**Copyright:** © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Clinical decision-making is a complex cognitive process in which healthcare professionals synthesize diverse information—such as patient symptoms, vital signs, medical history, and examination findings—to assess patient conditions and determine appropriate interventions [1]. The accuracy and timeliness of these decisions are critical for optimizing patient outcomes, particularly in high-pressure environments such as emergency departments (EDs) [2].

Triage is a fundamental component of ED operations, designed to prioritize patients based on the urgency of their medical condition [3]. Accurate triage ensures that critically ill patients receive timely attention while optimizing the allocation of limited resources for less urgent cases [3]. The Emergency Severity Index (ESI) is a widely adopted, standardized tool for assessing patient acuity and anticipating resource needs [4,5]. It categorizes patients

into five levels—ESI 1 (immediate), ESI 2 (emergency), ESI 3 (urgent), ESI 4 (semi-urgent), and ESI 5 (non-urgent)—based on symptoms, vital signs, and clinical judgment. Typically, trained nurses conduct triage assessments following structured protocols that incorporate information available at presentation, including chief complaints, high-risk symptoms, and vital signs. Despite these protocols, triage accuracy remains susceptible to clinician experience, cognitive bias, workload, and environmental pressures [6]. These challenges can lead to clinically significant errors, particularly under-triage, which may delay critical care, or over-triage, which can result in inefficient resource allocation [7].

Recent advances in generative artificial intelligence (AI), particularly large language models (LLMs), have demonstrated promising capabilities in processing unstructured clinical data and supporting decision-making tasks [1,8–11]. In the context of emergency care, LLMs have shown the ability to perform triage classification with performance approaching that of clinicians in controlled settings [12–17]. These findings suggest that LLMs may enhance triage workflows by improving consistency, supporting less experienced staff, and enabling scalable decision support in high-demand or remote environments [18]. However, despite these promising results, important limitations remain. LLM-based triage systems often rely on implicit reasoning learned during pretraining, without explicit grounding in structured clinical protocols. As a result, their predictions may be inconsistent, sensitive to prompt variations, and susceptible to hallucination. Critically, such limitations can introduce safety risks, including incorrect prioritization and under-triage of high-acuity patients. Furthermore, many existing approaches do not explicitly integrate domain-specific knowledge during inference, limiting their ability to produce clinically interpretable and guideline-consistent decisions.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address these limitations by incorporating external knowledge into the reasoning process of LLMs [19]. By retrieving relevant information at inference time and integrating it into the model's input, RAG can improve factual grounding, reduce hallucination, and enhance task-specific performance. In medical applications, prior studies have shown that augmenting LLMs with structured knowledge sources can improve diagnostic reasoning and triage accuracy [1,20,21]. However, existing RAG-based approaches in triage often rely on unstructured textual knowledge or large-scale corpora, with limited emphasis on encoding structured clinical decision rules that directly reflect triage protocols.

To address this gap, this study proposes the Triage Retrieval-Augmented Generation (TRAG) framework—a retrieval-augmented approach that dynamically integrates rule-based triage knowledge derived from ESI guidelines into LLM reasoning. In TRAG, triage knowledge is represented as "*if–then*" rules that capture clinically relevant decision criteria, enabling the model to align its predictions with established triage logic. By dynamically retrieving scenario-relevant rules and incorporating them into the model prompt, TRAG aims to improve the consistency, interpretability, and safety of triage predictions.

The objective of this study is to investigate whether integrating rule-based knowledge retrieval within a RAG framework can enhance LLM performance in predicting ESI levels from patient scenarios. Specifically, we evaluate whether TRAG improves classification accuracy, enhances agreement with ground-truth labels, and reduces clinically significant errors such as under-triage and over-triage, compared with non-augmented baseline models.

The remainder of this paper is organized as follows. Section 2 describes the TRAG framework, implementation details, and evaluation methodology. Section 3 presents experimental results across multiple models and retrieval settings. Section 4 discusses the findings, limitations, and clinical implications. Section 5 concludes the paper and outlines directions for future research.

## 2. Materials and Methods

This section describes the proposed Triage Retrieval-Augmented Generation (TRAG) framework, including its architecture, knowledge representation, implementation details, and evaluation methodology.

### *2.1. Triage-Based Retrieval-Augmented Generation Framework (TRAG)*

The Triage Retrieval-Augmented Generation (TRAG) framework integrates a large language model (LLM) with a dynamic rule-retrieval mechanism to enhance reasoning accuracy and interpretability in emergency triage prediction. This hybrid approach reduces model hallucination and ensures that triage predictions are firmly grounded in established clinical guidelines.

The TRAG framework consists of two primary components: the Retriever, which identifies the most relevant triage rules in response to an input query, and the Generator, which integrates the retrieved rules with the query to produce an informed classification. In this framework, the input is a patient scenario describing presenting symptoms, vital signs, medical history, and relevant clinical context. The output is a predicted ESI level (1–5). In addition, triage knowledge is represented as structured if–then rules that reflect clinical decision logic. Each rule follows the format:

IF [clinical condition(s)] THEN [ESI level]

For example: IF [a patient presents with severe respiratory distress and oxygen saturation below 90%], THEN [assign ESI Level 1].

As illustrated in Figure 1, the TRAG workflow consists of four main steps:

1. Query embedding: The patient scenario is encoded into a dense vector representation;
2. Semantic search: The query embedding is compared with stored rule embeddings using cosine similarity to find the most relevant rules;
3. Prompt augmentation: The top-k relevant rules are incorporated into the model prompt to provide contextual guidance;
4. Response generation: The LLM receives the patient scenario and retrieved rules and predicts the appropriate ESI level.

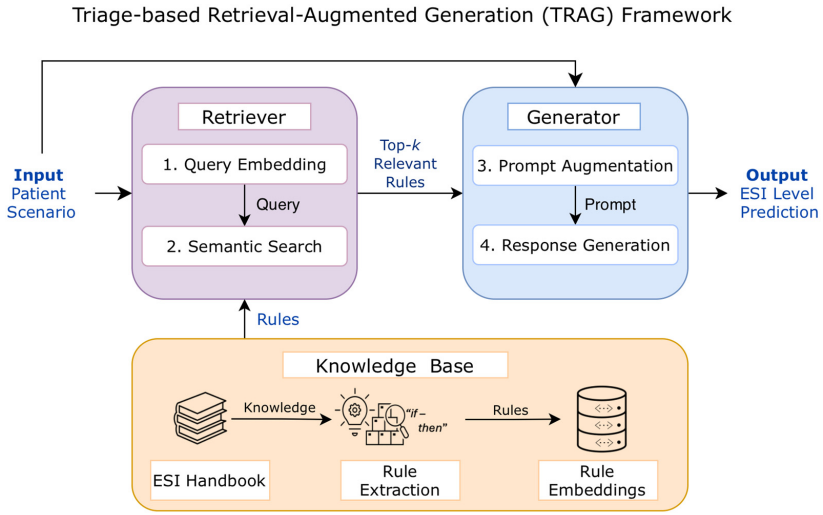



**Figure 1.** Workflow of the Triage Retrieval-Augmented Generation (TRAG) framework. (1) The patient scenario (query) is embedded into a dense vector representation, while ESI triage rules are stored as embeddings in the same vector space. (2) The retriever identifies the top-k most relevant rules. (3) The retrieved rules are then combined with the patient scenario to form an augmented prompt. (4) The prompt is provided to the large language model (LLM) to generate the predicted ESI level.

### 2.2. Implementation Details

#### 2.2.1. Retriever

The retrieval component was implemented using FAISS (Facebook AI Similarity Search) [22] to perform efficient similarity search in a high-dimensional embedding space. A flat index (IndexFlatIP) was used to enable exact nearest-neighbor retrieval. Embeddings for both patient scenarios and rule conditions were generated using OpenAI's text-embedding-3-large model [23]. To ensure consistent cosine similarity computation, L2 normalization was applied to both query embeddings and rule embeddings prior to indexing.

Given an input scenario, the retriever returns the Top-$k$ most semantically relevant rules, where $k \in \{5, 10, 15\}$. These values were selected to explore the trade-off between providing sufficient contextual guidance and avoiding excessive prompt complexity.

Semantic embedding was adopted to enable flexible alignment between patient descriptions and rule conditions expressed in natural language. This approach allows the system to handle variations in phrasing and partial matches that are common in clinical narratives.

#### 2.2.2. Generator

The generator component consists of a large language model (LLM) that produces the final ESI prediction. Three OpenAI LLMs were evaluated: GPT-4o-mini [24], GPT-4 [25], and GPT-3.5-Turbo [26]. All models were executed with a temperature of 0.2 to promote consistent and stable outputs while allowing limited variability. Other parameters were kept at their default API settings.

The model input consists of three components:

1. The patient scenario (query);
2. Retrieved triage rules (context);
3. Task instruction (ESI classification).

The model is instructed to output a single numeric label corresponding to the predicted ESI level (1–5).

#### 2.2.3. Knowledge Extraction

The knowledge base was derived from the *Emergency Severity Index (ESI) Handbook (Version 4)* [5], which provides standardized guidelines for triage decision-making. A total of approximately 120 triage rules were manually extracted by systematically identifying decision criteria described in the handbook and reformulating them into structured if–then rules. Each rule encodes clinically meaningful conditions, such as symptoms, vital signs, or high-risk indicators, mapped to a corresponding ESI level. To ensure clinical validity, the extracted rules were reviewed and verified by licensed physicians. While the rule set is not exhaustive, it captures key decision patterns defined in the ESI framework.

Each rule condition was embedded into a high-dimensional vector space to support semantic retrieval. Importantly, the rule-based knowledge base is independent of the evaluation dataset, ensuring no overlap or data leakage between training knowledge and test cases.

### 2.3. Performance Evaluation

The performance evaluation aims to assess whether integrating rule-based knowledge within the TRAG framework improves LLM performance in ESI classification. Two prompting settings—zero-shot and retrieval-augmented—were evaluated across multiple performance metrics. Specifically, the evaluation examines whether rule augmentation:

- Improves overall model performance compared with baseline (zero-shot setting);
- Reduces clinically significant errors such as under-triage and over-triage;
- Enhances agreement with ground-truth triage labels.

The evaluation dataset and prompt design are detailed below.

#### 2.3.1. Dataset

The evaluation dataset consists of 100 triage cases derived from the practice examples in *Chapter 9* of the *ESI Handbook* [5]. Each case is associated with a ground-truth ESI level (1–5). The class distribution is as follows: ESI 1 (16%), ESI 2 (27%), ESI 3 (26%), ESI 4 (15%), and ESI 5 (16%).

The original cases were reformulated into structured patient scenarios using GPT-4 to standardize input format while preserving clinically relevant information, including symptoms, vital signs, medications, and risk indicators. All reformulated cases were reviewed and validated by a licensed physician.

#### 2.3.2. Prompt Design

To evaluate the impact of contextual rule augmentation on triage classification, two prompting configurations were examined: zero-shot and retrieval-augmented settings.

In the zero-shot setting, the model receives only the patient scenario and is instructed to assign an ESI level (1–5). This serves as the baseline condition without external knowledge support.

In the retrieval-augmented setting (TRAG), the model receives additional contextual information derived from ESI guidelines. Specifically, the prompt incorporates (1) the patient scenario, (2) the top-$k$ retrieved triage rules, (3) structured ESI decision algorithm, and (4) task-specific instructions. This design guides the model to apply clinically grounded reasoning while maintaining consistency with the ESI triage algorithm.

The general ESI algorithm refers to the structured five-level triage decision process defined in the ESI handbook [5]. It first evaluates whether the patient requires immediate life-saving intervention (ESI Level 1) or presents with high-risk conditions, severe distress, or altered mental status (ESI Level 2). If neither applies, the triage level is determined based on the expected number of resources required (ESI Levels 3–5), with further adjustment based on abnormal vital signs when necessary.

The retrieval-augmented prompt is structured as follows:

- System role: Triage nurse assistant.
- Task: Assign an ESI level (1–5).
- Input: Patient scenario (clinical description), retrieved triage rules (top-$k$ relevant rules).
- ESI Algorithm:
  - A. Immediate life-saving intervention → ESI 1;
  - B. High-risk condition or severe distress → ESI 2;
  - C. Otherwise, assign ESI levels (3–5) based on expected resource needs;
  - D. Upgrade level if abnormal vital signs are present.
- Output format: JSON object containing the predicted ESI level.

This structured prompt enables the model to integrate scenario-specific information with general triage guidelines and retrieved rules during inference.

The following is a representative example of the input to the retrieval-augmented prompt with the resultant output.

Input:

- Patient scenario: "*A 4-day-old baby girl presents with a small amount of blood in the diaper. She is nursing well, has stable vital signs, and no additional symptoms.*"
- Retrieved rules ($k$ = 5):

1. IF *the patient is a neonate (<28 days old) with fever (≥38 °C)*, THEN *assign ESI Level 2*;

2. IF *the patient has upper respiratory infection symptoms and normal vital signs*, THEN *assign ESI Level 5*;
3. IF *the patient is pediatric (1–3 months old) with fever >38 °C*, THEN *consider ESI Level 2 and assess appearance*;
4. IF *the patient presents with petechial rash and altered mental status*, THEN *assign ESI Level 1*;
5. IF *the patient has mild symptoms and is active with normal intake*, THEN *assign ESI Level 5*.

Output: {"level": "5"}.

#### 2.3.3. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and Quadratic Weighted Kappa (QWK).

Macro-averaged precision, recall, and F1-score were used to ensure equal weighting across all ESI classes. Given the ordinal nature of ESI levels, QWK was used to measure agreement between predicted and ground-truth ESI levels while accounting for the ordinal nature of the classes. QWK is defined as:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}$$

where $O_{ij}$ and $E_{ij}$ represent the observed and expected agreement matrices, respectively, and $w_{ij}$ is the quadratic weight defined as:

$$w_{ij} = \frac{(i-j)^2}{(k-1)^2}$$

where $i$ and $j$ are the predicted and true class indices, and $k$ is the number of categories. This formulation penalizes larger disagreements more heavily, making it suitable for ordinal triage classification.

Predictions were further categorized as correct triage, over-triage, or under-triage. Correct triage indicates exact agreement between predicted and true ESI levels. Over-triage occurs when the model assigns a higher (more urgent) level than appropriate, whereas under-triage reflects the assignment of a lower (less urgent) level. The proportion of each category was calculated. These categories were analyzed separately due to their clinical significance, particularly the safety risks associated with under-triage.

Finally, a confusion matrix was constructed to analyze performance across ESI levels.

All experiments were conducted using a deterministic configuration (*temperature* = 0.2), and results are reported from a single run.

## 3. Results

This section presents the experimental evaluation of the TRAG framework using three large language models (LLMs)—GPT-3.5, GPT-4, and GPT-4o-mini—tested under multiple TRAG settings with varying numbers of retrieved rules ($k$ = 5, 10, and 15) and compared against their respective zero-shot baselines.

### *3.1. Overall Model Performance*

The overall performance of the evaluated models across different settings is summarized in Table 1. For GPT-3.5, the zero-shot setting yielded modest results (accuracy 0.45, F1-score 0.332, and QWK 0.668). Performance improved across multiple metrics under TRAG settings (illustrated in Figure 2), with the Top-15 achieving the highest metrics (accuracy 0.68, F1-score 0.682, QWK 0.820), indicating that retrieval augmentation improved the model's ability to identify ESI levels correctly.

**Table 1.** Performance metrics for LLMs under zero-shot and TRAG settings. Accuracy, precision, recall, F1-score, and Quadratic Weighted Kappa (QWK) are reported for GPT-3.5, GPT-4, and GPT-4o-mini across zero-shot and TRAG (Top-*k* = 5, 10, 15) settings.

| Model | Setting | Accuracy | Precision | Recall | F1-Score | QWK |
|---|---|---|---|---|---|---|
| GPT 3.5 | Zero-shot | 0.45 | 0.649 | 0.377 | 0.332 | 0.668 |
| | TRAG Top-5 | 0.66 | 0.765 | 0.641 | 0.672 | 0.804 |
| | TRAG Top-10 | 0.66 | 0.735 | 0.64 | 0.661 | 0.815 |
| | TRAG Top-15 | 0.68 | 0.754 | 0.661 | 0.682 | 0.82 |
| GPT 4 | Zero-shot | 0.7 | 0.714 | 0.697 | 0.7 | 0.868 |
| | TRAG Top-5 | 0.73 | 0.782 | 0.707 | 0.727 | 0.816 |
| | TRAG Top-10 | 0.76 | 0.8 | 0.728 | 0.746 | 0.819 |
| | TRAG Top-15 | 0.76 | 0.798 | 0.739 | 0.756 | 0.847 |
| GPT 4o-mini | Zero-shot | 0.68 | 0.816 | 0.653 | 0.675 | 0.795 |
| | TRAG Top-5 | 0.76 | 0.776 | 0.742 | 0.751 | 0.826 |
| | TRAG Top-10 | 0.76 | 0.776 | 0.742 | 0.751 | 0.826 |
| | TRAG Top-15 | 0.75 | 0.77 | 0.735 | 0.743 | 0.824 |

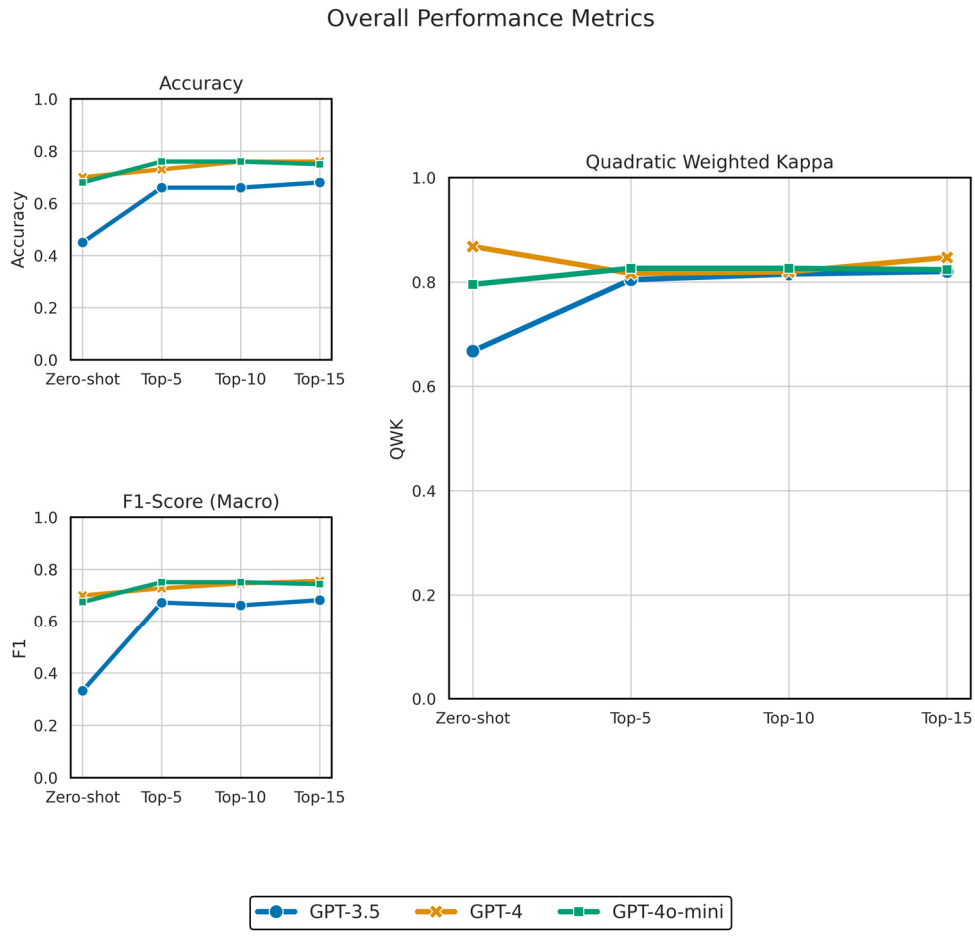



**Figure 2.** Overall performance of LLMs in ESI prediction across zero-shot and TRAG settings. Performance is reported in terms of accuracy, F1-score, and Quadratic Weighted Kappa (QWK) for GPT-3.5, GPT-4, and GPT-4o-mini under zero-shot and TRAG (Top-*k* = 5, 10, 15) settings. Each panel illustrates the variation in performance metrics as the number of retrieved rules increases.

GPT-4 exhibited stronger baseline performance in the zero-shot setting (accuracy 0.70, F1-score 0.700, QWK 0.868), which was further improved under TRAG. The Top-10 and Top-15 settings achieved the highest accuracy (0.76) and F1-score (0.756), with QWK reaching 0.847.

GPT-4o-mini achieved a zero-shot accuracy of 0.68 and an F1-score of 0.675. Incorporating the TRAG framework consistently enhanced performance, with Top-5 and Top-10 retrieval settings reaching accuracies of 0.76 and F1-scores of 0.751, and QWK values ranging from 0.824 to 0.826.

Statistical significance was evaluated using McNemar's test. For GPT-3.5, the improvements observed under TRAG were statistically significant (e.g., $p = 0.0004$ for Top-15), with substantially more correct than incorrect predictions. For GPT-4 and GPT-4o-mini, although improvements were observed, these differences were not statistically significant ($p > 0.05$).

### 3.2. Under-Triage and Over-Triage Analysis

The distribution of triage predictions across models and settings is shown in Figure 3.

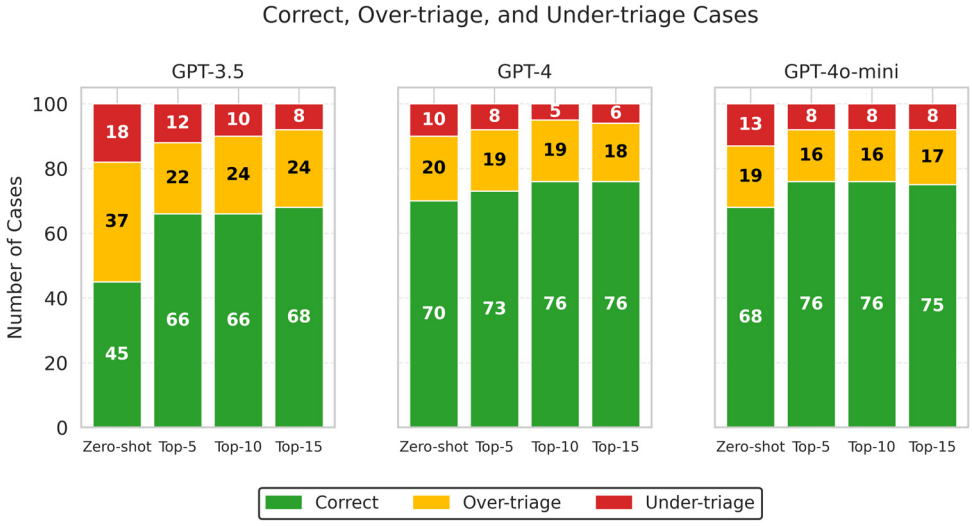



**Figure 3.** Distribution of triage outcomes across models and settings. Predictions are categorized as correct triage, over-triage, or under-triage for GPT-3.5, GPT-4, and GPT-4o-mini under zero-shot and TRAG (Top-$k$ = 5, 10, 15) settings.

For GPT-3.5, the zero-shot setting yielded 45% correct triage predictions, with 37% over-triaged and 18% under-triaged cases. Incorporating TRAG markedly improved correct predictions (68% under Top-15), while over-triage and under-triage rates decreased to 24% and 8%, respectively. GPT-4 demonstrated higher baseline performance, with 70% correct predictions in the zero-shot setting. TRAG further improved correct triage rates, achieving up to 76% under Top-10 and Top-15 settings, and reduced under-triage rates to as low as 5%. Over-triage remained stable (18–19%). GPT-4o-mini followed similar trends, with zero-shot correct predictions of 68% increasing to 76% under Top-5 and Top-10 TRAG settings. Over-triage decreased from 19% to 16%, while under-triage reduced from 13% to 8%.

Overall, TRAG increased the proportion of correct triage predictions while reducing both over- and under-triage across all models in most settings.

### 3.3. Performance by ESI Level

As shown in Figure 4, the confusion matrices illustrate model performance across ESI levels under zero-shot and TRAG settings. GPT-3.5 in the zero-shot condition failed to reliably identify critical ESI-1 patients (1/16 correct), but TRAG substantially enhanced recognition to 10–13 correct cases, demonstrating a clear reduction in under-triage. Recog-

nition of ESI-2 cases remained high (24–25 correct cases), while predictions for stable patients (ESI-4–5) showed notable improvement, particularly for ESI-5 (increasing from 1 to 10–12 correct cases).

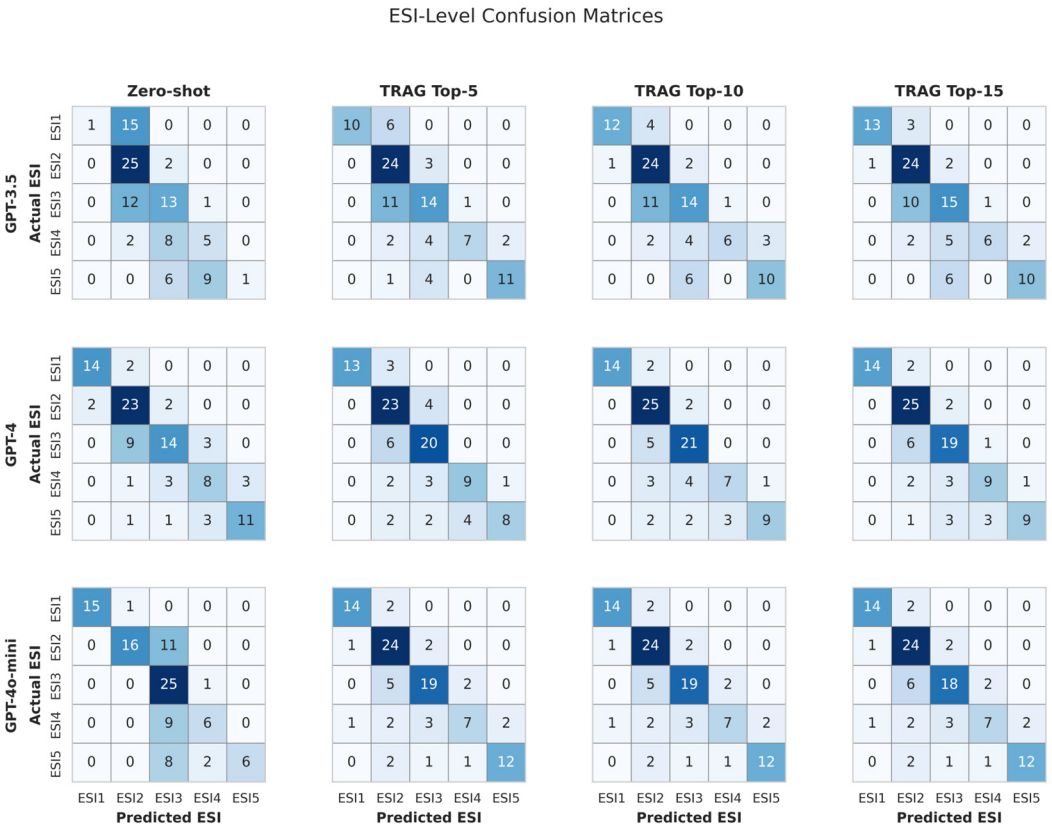



**Figure 4.** Confusion matrices for ESI classification across models and TRAG settings. Each row corresponds to a model (GPT-3.5, GPT-4, GPT-4o-mini), and each column represents a configuration (zero-shot, TRAG Top-*k* = 5, 10, 15). Cell values indicate the number of cases for each true versus predicted ESI level, with darker shades representing higher counts. These matrices provide detailed insight into class-wise performance and misclassification patterns.

GPT-4 showed high baseline accuracy for critical cases (ESI-1-2), with modest gains under TRAG. Mid-acuity cases (ESI-3) showed slight improvement, while stable cases (ESI-4–5) remained unchanged. GPT-4o-mini followed similar patterns, with TRAG enhancing recognition of both critical and low-acuity cases—particularly ESI-2 and ESI-5—while mid-acuity classifications (ESI 3–4) remained stable.

Overall, these results indicate that TRAG enhanced the recognition of both critical and low-acuity patients, particularly in models with lower baseline accuracy, while maintaining stable performance in high-performing models.

To further examine whether performance improvements were driven by clinically critical cases, per-class metrics were analyzed for high-acuity levels (ESI-1 and ESI-2). For GPT-3.5, TRAG substantially improved ESI-1 recall from 0.062 in the zero-shot setting to 0.812 under the Top-15 configuration, indicating a marked reduction in missed critical cases. For GPT-4, ESI-1 recall remained stable (0.875), while ESI-2 recall improved from 0.852 to 0.926. Similarly, GPT-4o-mini showed improved ESI-2 recall from 0.593 to 0.889, although ESI-1 recall decreased slightly while remaining high overall. These results indicate that the observed performance gains are not solely driven by low-acuity cases but also reflect improved detection of high-acuity patients in several configurations.

## 4. Discussion

This study evaluated the impact of integrating rule-based knowledge within a retrieval-augmented generation framework (TRAG) for emergency triage prediction. The results indicate that incorporating structured ESI-based rules can improve model performance in several settings, particularly for models with lower baseline capability.

Across all evaluated models, retrieval augmentation generally increased classification accuracy and macro F1-scores. The most pronounced improvements were observed for GPT-3.5, which showed substantial gains in both overall performance and agreement with ground-truth labels. These improvements were statistically significant based on McNemar's test ($p < 0.001$). In contrast, higher-performing models such as GPT-4 exhibited more modest and less consistent improvements. Notably, agreement as measured by QWK did not uniformly improve across all TRAG configurations, with some settings showing slight decreases relative to the zero-shot baseline. These findings suggest that the benefits of retrieval augmentation are model-dependent and may be influenced by the interaction between model capability and external knowledge integration.

Analysis of triage outcomes provides additional insight into model behavior. TRAG increased the proportion of correct classifications in most settings and was associated with reductions in under-triage for lower-performing models. This is particularly relevant from a clinical perspective, as under-triage represents a higher-risk error due to potential delays in critical care. However, reductions in under-triage were not consistently observed across all models and settings, and over-triage rates remained relatively stable. These results indicate that while retrieval-augmented reasoning can support safer decision patterns in some cases, its impact on clinical risk is not uniform.

At the level of individual ESI categories, TRAG contributed to improved recognition of both high-acuity (ESI 1–2) and low-acuity (ESI 5) cases, particularly for models with weaker baseline performance. In contrast, predictions for mid-acuity cases (ESI 3–4) remained largely unchanged. This pattern suggests that rule-based augmentation may be more effective in scenarios where decision boundaries are more clearly defined by explicit criteria, while offering limited benefit in more ambiguous cases. Moreover, improvements were observed in high-acuity categories (ESI-1 and ESI-2), particularly in recall, indicating that TRAG can enhance the detection of critical cases rather than solely improving performance on low-acuity classifications.

An important observation is that performance did not increase monotonically with larger values of $k$ (number of retrieved rules). In some cases, increasing $k$ resulted in marginal gains or slight performance degradation. This may reflect a trade-off between providing relevant contextual guidance and introducing additional noise or redundancy into the prompt. These findings highlight the importance of retrieval quality and prompt design in determining the effectiveness of retrieval-augmented approaches.

Overall, a clear improvement is associated with the integration of structured, clinically grounded knowledge into the model's reasoning process. By retrieving scenario-relevant rules and incorporating them into the prompt, TRAG provides structured guidance that can reduce ambiguity and support more consistent decision-making. At the same time, the variability in results indicates that the effectiveness of this approach depends on both the quality of retrieved rules and the underlying model's ability to utilize them effectively.

Several limitations should be acknowledged. First, the evaluation dataset consisted of 100 cases derived from handbook examples, which may not fully capture the diversity and complexity of real-world emergency presentations. Second, experiments were conducted using a single deterministic configuration without repeated trials, limiting the ability to draw definitive conclusions about performance differences. Third, the rule set, while clinically validated, is not exhaustive and may not cover all triage scenarios. Additionally,

the retrieval component was not independently evaluated, making it difficult to assess retrieval quality in isolation. Despite these limitations, the findings provide preliminary evidence that integrating structured rule-based knowledge within a retrieval-augmented framework can enhance LLM-based triage prediction, particularly for models with lower baseline performance. The TRAG framework demonstrates the potential of combining structured clinical logic with generative reasoning to support more transparent and context-aware decision-making.

## 5. Conclusions

This study presented the Triage Retrieval-Augmented Generation (TRAG) framework, a domain-specific approach that integrates structured, rule-based clinical knowledge with large language model (LLM) reasoning for emergency triage prediction. By incorporating scenario-relevant ESI rules into the inference process, TRAG aims to improve the consistency and clinical grounding of LLM-based triage decisions.

Experimental results indicate that retrieval-augmented prompting improves classification performance in several settings, particularly for models with lower baseline capability. Notably, improvements were observed in overall accuracy, agreement with ground-truth labels, and reductions in under-triage in selected configurations. For higher-performing models, gains were more modest and dependent on retrieval settings, suggesting that the effectiveness of rule-based augmentation varies with model characteristics.

Future work will focus on expanding the rule base, improving retrieval quality, incorporating retrieval-specific evaluation metrics, and validating the framework on larger and more diverse clinical datasets. Such extensions are necessary to further assess the robustness and practical applicability of retrieval-augmented approaches in real-world clinical settings. In addition, comparison with traditional machine learning models trained on the same knowledge base will be explored to provide a more comprehensive evaluation.

**Funding:** This research was funded by Ongoing Research Funding program, (ORF-2026-1983), King Saud University, Riyadh, Saudi Arabia.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset used in this study was derived from publicly available examples in the *Emergency Severity Index (ESI) Handbook, Version 4*, and reformulated for research purposes. The processed dataset is available from the corresponding author upon reasonable request.

**Acknowledgments:** The author gratefully acknowledges the physicians who reviewed the triage dataset and provided valuable clinical guidance supporting this research.

**Conflicts of Interest:** The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ESI | Emergency Severity Index |
| FAISS | Facebook AI Similarity Search |
| LLM | Large Language Model |
| QWK | Quadratic Weighted Kappa |
| RAG | Retrieval-Augmented Generation |
| TRAG | Triage Retrieval-Augmented Generation |

## References


1. Gaber, F.; Shaik, M.; Allega, F.; Bilecz, A.J.; Busch, F.; Goon, K.; Franke, V.; Akalin, A. Evaluating Large Language Model Workflows in Clinical Decision Support for Triage and Referral and Diagnosis. *npj Digit. Med.* **2025**, *8*, 263. [CrossRef] [PubMed]
2. Sutriningsih, A.; Wahyuni, C.U.; Haksama, S. Factors Affecting Emergency Nurses' Perceptions of the Triage Systems. *J. Public Health Res.* **2020**, *9*, 1808. [CrossRef] [PubMed]
3. Kaboudi, N.; Firouzbakht, S.; Shahir Eftekhar, M.; Fayazbakhsh, F.; Joharivarnoosfaderani, N.; Ghaderi, S.; Dehdashti, M.; Mohtasham Kia, Y.; Afshari, M.; Vasaghi-Gharamaleki, M.; et al. Diagnostic Accuracy of ChatGPT for Patients' Triage; a Systematic Review and Meta-Analysis. *Arch. Acad. Emerg. Med.* **2024**, *12*, e60. [CrossRef] [PubMed]
4. Badashova, K.; Shesser, R. Chapter 6—Triage. In *The Emergency Department Technician Handbook*; Shesser, R., Pourmand, A., Keim, A., Eds.; Elsevier: New Delhi, India, 2024; pp. 38–42, ISBN 978-0-323-83002-7.
5. Gilboy, N.; Tanabe, T.; Travers, D.; Rosenau, A.M. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4*; Implementation Handbook 2012 Edition; Agency for Healthcare Research and Quality: Rockville, MD, USA, 2011.
6. Chmielewski, N.; Moretz, J. ESI Triage Distribution in U.S. Emergency Departments. *Adv. Emerg. Nurs. J.* **2022**, *44*, 46–53. [CrossRef] [PubMed]
7. Hitchcock, M.; Gillespie, B.; Crilly, J.; Chaboyer, W. Triage: An Investigation of the Process and Potential Vulnerabilities. *J. Adv. Nurs.* **2014**, *70*, 1532–1541. [CrossRef] [PubMed]
8. Abbas, A.; Rehman, M.S.; Rehman, S.S. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus* **2024**, *16*, e55991. [CrossRef] [PubMed]
9. Ayoub, M.; Ballout, A.A.; Zayek, R.A.; Ayoub, N.F. Mind + Machine: ChatGPT as a Basic Clinical Decisions Support Tool. *Cureus* **2023**, *15*, e43690. [CrossRef] [PubMed]
10. Jiang, L.Y.; Liu, X.C.; Nejatian, N.P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H.A.; Laufer, I.; Punjabi, P.; et al. Health System-Scale Language Models Are All-Purpose Prediction Engines. *Nature* **2023**, *619*, 357–362. [CrossRef] [PubMed]
11. Park, Y.-J.; Pillai, A.; Deng, J.; Guo, E.; Gupta, M.; Paget, M.; Naugler, C. Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 72. [CrossRef] [PubMed]
12. Ito, N.; Kadomatsu, S.; Fujisawa, M.; Fukaguchi, K.; Ishizawa, R.; Kanda, N.; Kasugai, D.; Nakajima, M.; Goto, T.; Tsugawa, Y. The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study. *JMIR Med. Educ.* **2023**, *9*, e47532. [CrossRef] [PubMed]
13. Arslan, B.; Nuhoglu, C.; Satici, M.O.; Altinbilek, E. Evaluating LLM-Based Generative AI Tools in Emergency Triage: A Comparative Study of ChatGPT Plus, Copilot Pro, and Triage Nurses. *Am. J. Emerg. Med.* **2025**, *89*, 174–181. [CrossRef] [PubMed]
14. 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*, 146–150. [CrossRef] [PubMed]
15. Meyer, N.S.; Meyer, J.W. A Practical Guide to the Utilization of ChatGPT in the Emergency Department: A Systematic Review of Current Applications, Future Directions, and Limitations. *Cureus* **2025**, *17*, e81802. [CrossRef] [PubMed]
16. Paslı, S.; Şahin, A.S.; Beşer, M.F.; Topçuoğlu, H.; Yadigaroğlu, M.; İmamoğlu, M. Assessing the Precision of Artificial Intelligence in ED Triage Decisions: Insights from a Study with ChatGPT. *Am. J. Emerg. Med.* **2024**, *78*, 170–175. [CrossRef] [PubMed]
17. Williams, C.Y.K.; Zack, T.; Miao, B.Y.; Sushil, M.; Wang, M.; Kornblith, A.E.; Butte, A.J. Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. *JAMA Netw. Open* **2024**, *7*, e248895. [CrossRef] [PubMed]
18. Lee, S.; Jung, S.; Park, J.-H.; Cho, H.; Moon, S.; Ahn, S. Performance of ChatGPT, Gemini and DeepSeek for Non-Critical Triage Support Using Real-World Conversations in Emergency Department. *BMC Emerg. Med.* **2025**, *25*, 176. [CrossRef] [PubMed]
19. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997.
20. Yazaki, M.; Maki, S.; Furuya, T.; Inoue, K.; Nagai, K.; Nagashima, Y.; Maruyama, J.; Toki, Y.; Kitagawa, K.; Iwata, S.; et al. Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model. *Prehosp. Emerg. Care* **2025**, *29*, 203–209. [CrossRef] [PubMed]
21. Gan, R.K.; Uddin, H.; Gan, A.Z.; Yew, Y.Y.; González, P.A. ChatGPT's Performance before and after Teaching in Mass Casualty Incident Triage. *Sci. Rep.* **2023**, *13*, 20350. [CrossRef] [PubMed]
22. Ghadekar, P.P.; Mohite, S.; More, O.; Patil, P.; Sayantika; Mangrule, S. Sentence Meaning Similarity Detector Using FAISS. In Proceedings of the 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 18–19 August 2023; pp. 1–6.
23. OpenAI. Text-Embedding-3-Large. Available online: https://developers.openai.com/api/docs/models/text-embedding-3-large (accessed on 3 November 2025).
24. OpenAI. GPT-4o-Mini. Available online: https://platform.openai.com/docs/models/gpt-4o-mini (accessed on 4 November 2025).

25. OpenAI. GPT-4. Available online: https://platform.openai.com/docs/models/gpt-4 (accessed on 4 November 2025).
26. OpenAI. GPT-3.5-Turbo. Available online: https://platform.openai.com/docs/models/gpt-3.5-turbo (accessed on 4 November 2025).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.