SOAPTriage: SOAP-Guided Multi-View Clinical Text Modeling Framework for Automated ESI Prediction

Enming Wang[1,5,*], Jianlei Wang[1,5,*] Xueping Peng[2], Hongjiao Guan[1,5], Yinglong Wang[1,5], Sibo Wei[3], Jianbin Guo[3], Ruifeng Xu[4], Wenpeng Lu[1,5,†]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2]Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia

[3]Beijing Wenge Technology Co., Ltd., Beijing, China [4]Harbin Institute of Technology, Shenzhen, China

[5]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China

enming.wang.2025@foxmail.com, jianlei.wang.qlu@gmail.com, wenpeng.lu@qlu.edu.cn

Abstract

Emergency departments (ED) rely on the Emergency Severity Index (ESI) to assess patient acuity and prioritize care, a process that is largely driven by clinical triage text. Despite recent progress in automated ESI prediction, two fundamental challenges remain: the scarcity of high-quality triage text data due to privacy and regulatory constraints and the lack of a clinically grounded triage framework capable of explicitly capturing the multidimensional structure of triage reasoning. To address these challenges, we draw inspiration from the clinically grounded SOAP paradigm, in which SOAP refers to Subjective, Objective, Assessment, and Plan and captures four complementary aspects of clinical reasoning. Building on this paradigm, we propose SOAPTriage, a SOAP-guided multi-view clinical text modeling framework for automated ESI prediction. To mitigate data scarcity, SOAPTriage introduces a Clinical Note Augmentation (CNA) module that generates natural-language triage notes from structured ED records, resulting in 15,393 augmented clinical notes derived from a real-world dataset. To incorporate clinical structure, SOAPTriage employs a SOAP-Guided Encoding (SGE) module that models patient conditions from four complementary SOAP perspectives, together with an adaptive SOAP-Aware Aggregation and Inference (SAAI) module that performs multi-view reasoning to infer ESI levels. Extensive experiments show that SOAPTriage consistently outperforms strong prompting-based, multi-agent, and encoder-based baselines, demonstrating the effectiveness of SOAP-guided multi-view clinical text modeling for automated emergency triage.[1]

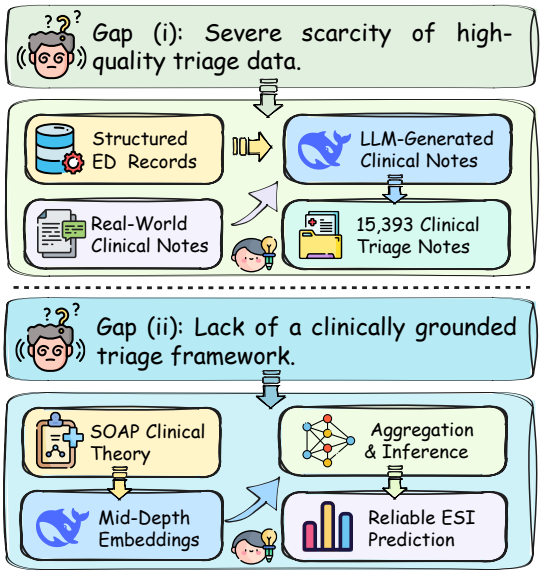

Figure 1: Overview of SOAPTriage. The framework tackles triage data scarcity by generating clinical notes from structured records and real-world notes. It also addresses the lack of a clinically grounded triage framework by incorporating SOAP-guided reasoning, extracting multi-view representations, and aggregating them to infer reliable ESI predictions.

1 Introduction

The Emergency Severity Index (ESI) is a crucial clinical triage indicator that is widely adopted in emergency departments (ED) (Scherer et al., 2017; Williams et al., 2024a). It provides a standardized and structured framework to assess the urgency and severity of patients' conditions, ranging from critical (level 1) to non-urgent cases (level 5), enabling clinicians to prioritize treatment and allocate medical resources effectively (Lu et al., 2024). Accurate ESI assignment is therefore essential to ensure timely care delivery, efficient resource utilization, and patient safety.

[*] Equal contribution

[†] Corresponding author

[1]Our code and datasets can be found at https://github.com/xiaoyaoiii/SOAPTriage.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34813–34837

Despite its importance, ESI assignment still relies heavily on manual expert judgment through the review of clinical notes, making the process time-consuming and labor-intensive. As patient volumes continue to rise, this manual assignment inevitably leads to clinician fatigue and reduced consistency in triage decisions, which in turn can cause resource misallocation, delays in critical care, and decreased patient satisfaction (Da'Costa et al., 2025). These challenges are particularly pronounced during peak ED hours, underscoring the urgent need for accurate and reliable automated ESI prediction systems to support ED triage (Shen et al., 2025; Lu et al., 2025).

A growing body of research has explored automated ESI prediction using machine learning and, more recently, large language models (LLMs). While machine learning-based methods can assist in clinical triage, they typically rely on structured inputs and manually engineered features, which limits their scalability and robustness in dynamic and complex real-world clinical settings (Joseph et al., 2020; Yao et al., 2021; Sánchez-Salmerón et al., 2022). More recent LLM-based methods have demonstrated strong potential for clinical reasoning. Their applications to triage include directly inferring triage levels from free-text complaints and clinical histories (Law et al., 2024; Levine et al., 2024; Shen et al., 2025), as well as multi-agent frameworks that decompose the triage process into specialized sub-tasks (Lu et al., 2024). Despite these advances of LLM-based approaches, building accurate and reliable automated ESI prediction systems for real-world ED environments remains a significant challenge.

We identify two major gaps that hinder the practical development of LLM-based ESI prediction systems. **Gap 1: Severe scarcity of high-quality triage data**. Due to privacy protections and regulatory constraints, real clinical triage notes are difficult to release publicly. As a result, most existing studies rely on relatively small or quality-inconsistent datasets, which fail to capture the diversity and complexity of real-world triage scenarios. **Gap 2: Lack of a clinically grounded triage framework.** Many existing approaches focus primarily on model architectures, prompt engineering, or local feature modeling, while neglecting to explicitly integrate systematic clinical triage theories into the inference process. Consequently, these models may not comprehensively account for patients' multidimensional clinical information, leading to inaccurate ESI predictions.

To bridge these gaps, we propose SOAPTriage, a SOAP-guided multi-view clinical text modeling framework for automated ESI prediction. To address **Gap 1**, SOAPTriage introduces a *Clinical Note Augmentation* (CNA) module that employs a template-based generation strategy to transform routine structured emergency department (ED) records into natural-language clinical triage notes. Using CNA, we construct a large-scale triage dataset by generating 15,393 clinical notes from visits in the MIMIC-IV dataset (Johnson et al., 2020), providing a solid data foundation for automated triage research. To address **Gap 2**, SOAPTriage explicitly incorporates the clinically grounded SOAP (Subjective, Objective, Assessment, Plan) triage paradigm (Crausman, 1998; Jaroudi and Payne, 2019) as an inductive bias in model inference. Specifically, we design a *SOAP-Guided Encoding* (SGE) module that models patient conditions from four complementary clinical perspectives and produces multi-view representations. These representations are further integrated by an adaptive *SOAP-Aware Aggregation and Inference* (SAAI) module, which captures clinicians' holistic reasoning across multiple clinical dimensions to predict the final ESI level. Main contributions of this work are summarized below:

- We propose a novel Clinical Note Augmentation (CNA) method that enables the construction of a natural-language ED triage dataset. Using CNA, we generate 15,393 clinical triage notes with ESI labels from MIMIC-IV, substantially alleviating data scarcity in automated ESI prediction research.
- We introduce a SOAP-guided multi-view clinical text modeling framework for automated ESI prediction. Through the proposed SGE and SAAI modules, SOAPTriage explicitly incorporates clinically grounded SOAP theory to model patient conditions from multiple complementary perspectives and perform adaptive multi-view reasoning.
- Extensive experiments demonstrate that SOAPTriage consistently outperforms strong prompting-based, multi-agent, and encoder-based baselines, highlighting the effectiveness of integrating clinically grounded SOAP priors into automated ESI prediction.

2 Related Work

2.1 Data Augmentation

Data augmentation (DA) aims to alleviate data scarcity by generating additional training samples from existing data, thereby improving model generalization (Rentschler et al., 2022; Dai et al., 2025; Ma et al., 2025; Wang et al., 2026). Existing DA methods for text can be broadly categorized into transformation-based, prompt-driven generative, and retrieval-based approaches.

Transformation-based augmentation creates new training samples by applying small surface-level edits to existing instances, such as synonym replacement, insertion, deletion, or word reordering (Wei and Zou, 2019; Guo et al., 2022). These methods are simple, but their conservative transformations limit semantic diversity and the extent of data space expansion. *Prompt-driven generative augmentation* generates additional training samples by prompting large language models to synthesize new instances conditioned on task descriptions or existing examples (Dai et al., 2025; Lee et al., 2024; Zhang et al., 2025b). While this approach can increase sample diversity, it is sensitive to prompt design and may produce instances with inconsistent quality. In contrast, *retrieval-based augmentation* enriches model inputs by retrieving relevant information from external corpora or knowledge bases and incorporating it into the generation or decision process, thereby improving factual consistency (Yang et al., 2024; Zhan et al., 2025).

In this work, we focus on leveraging structured ED records together with retrieved real-world clinical documentation to guide LLMs in generating natural-language clinical triage notes for data augmentation.

2.2 Automated Triage

Automated triage aims to leverage artificial intelligence and machine learning (ML) to support ESI assessment by evaluating the urgency and severity of patients' conditions (Friedman et al., 2024; El Arab and Al Moosa, 2025). Existing research can be broadly categorized into three directions: traditional supervised ML triage approaches, language model-based triage approaches, and multi-agent triage approaches.

Traditional supervised ML triage approaches apply supervised learning models to structured clinical variables, such as vital signs, to predict ESI levels (Joseph et al., 2020; Yao et al., 2021; Karlafti et al., 2023). These methods rely heavily on annotated datasets and structured inputs, which limits their flexibility and robustness in dynamic, real-world emergency department environments. *Language model-based triage approaches* leverage the clinical knowledge embedded in LLMs through prompt-based inference or task-specific pretraining and fine-tuning to perform diagnosis and triage directly from natural-language clinical descriptions (Williams et al., 2024b; Levine et al., 2024; Zhang et al., 2025a; Xu et al., 2025; Wang et al., 2025). Meanwhile, *multi-agent triage approaches* simulate collaborative decision-making by decomposing the triage process into interacting agents or components, often integrating retrieval-augmented generation and hierarchical ESI prediction (Lu et al., 2024). These methods highlight the potential of LLMs for automated triage, but they primarily focus on model architectures, prompt engineering, or local feature modeling, while neglecting to explicitly integrate systematic clinical triage theories into the inference process.

Building on this gap, our work focuses on incorporating the clinically grounded SOAP triage paradigm into the model inference process, enabling more reliable and clinically consistent ESI prediction.

3 Methodology

As shown in Figure 2, SOAPTriage is a unified framework for automated ESI prediction, consisting of three main modules: *Clinical Note Augmentation* (CNA), *SOAP-Guided Encoding* (SGE), and *SOAP-Aware Aggregation and Inference* (SAAI). First, given a structured ED record, the CNA module converts structured clinical fields into natural-language triage notes. It enables large-scale data construction by transforming privacy-preserving structured records into realistic clinical notes while retaining clinically relevant information. Next, the SGE module takes the augmented triage notes as input and extracts multiple SOAP-aligned representations. Specifically, it follows the clinically established SOAP paradigm, guiding a frozen LLM to encode each triage note from four complementary clinical perspectives: Subjective (S), Objective (O), Assessment (A), and Plan (P), yielding SOAP-guided mid-depth representations. These representations are then passed to the SAAI module to predict the final ESI level. Rather than treating all SOAP components equally, SAAI learns adap-

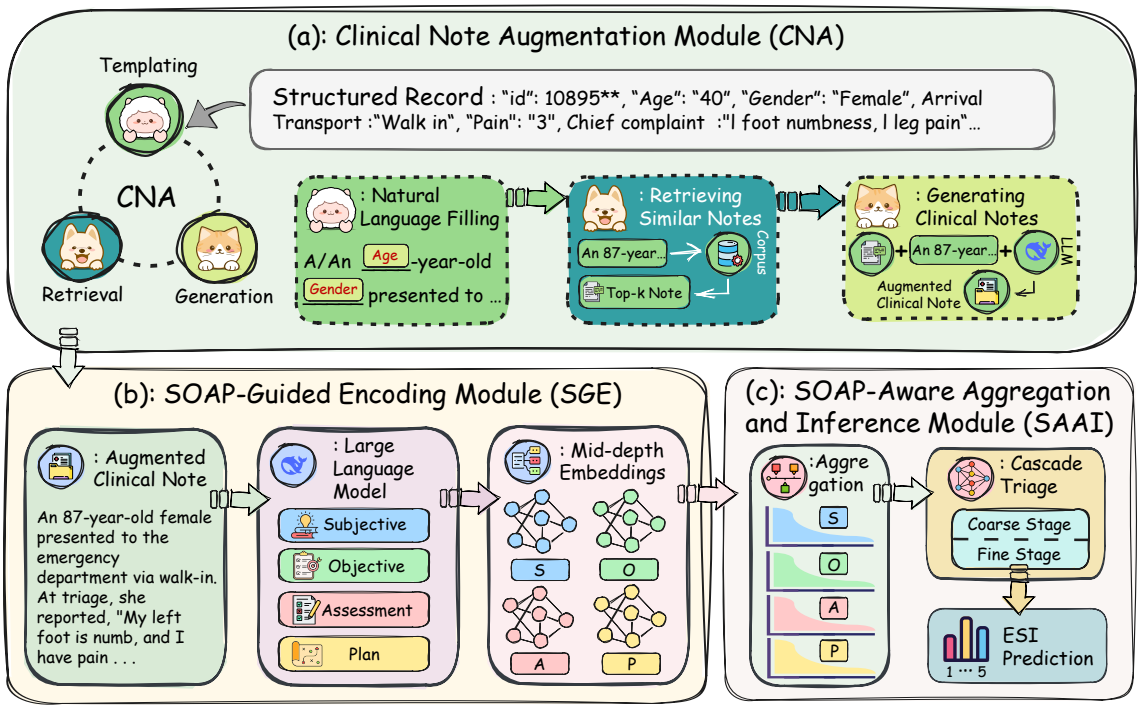

Figure 2: Architecture of SOAPTriage. The CNA module converts structured ED records into natural-language triage notes to construct a high-quality dataset with 15,393 samples. The SGE module, guided by SOAP triage theory, comprehensively assesses patients' conditions and extracts mid-depth representations from four perspectives. The SAAI module adaptively aggregates these SOAP-aware representations to infer the final ESI level.

tive importance weights over the SOAP representations and integrates them accordingly, reflecting how clinicians emphasize different clinical signals based on the triage context and perceived risk. We next formalize the ESI prediction task and describe each module of SOAPTriage in detail.

3.1 Task Definition

Given a natural-language triage note describing a patient's ED visit, the automated ESI prediction task aims to assess the urgency and severity of the patient's conditions to predict an ESI level, which is used to prioritize treatment and allocate resources. Formally, let $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ denote a collection of clinical notes, where each note v_i is a natural-language note corresponding to an ED visit, integrating multiple clinical and contextual information (e.g., chief complaint, vital signs, arrival time, and visit context). The objective is to learn a model that maps each note v_i to a corresponding triage ESI label $y_i \in \{1, \cdots, 5\}$, where lower values indicate higher clinical urgency.

3.2 Clinical Note Augmentation Module

Due to privacy protections and regulatory constraints, high-quality triage notes expressed in free-form natural language are extremely scarce. However, medical institutions maintain a large volume of routinely collected structured ED records. To address this data scarcity, we design a Clinical Note Augmentation (CNA) module that converts structured ED records into natural-language triage notes. As shown in Figure 2(a), CNA is composed of three sequential components: *Templating (natural language filling)*, *Retrieval (retrieving similar notes)*, and *Generation (generating clinical notes)*. Working together, these components transform structured ED records into clinical triage notes, thereby establishing a robust data foundation for the ESI prediction task.

Templating. ED triage records are primarily stored as structured clinical fields, whereas real-world ED documentation is typically expressed in free-form natural language (Lin et al., 2024). To bridge this representational gap, the *Templating* component converts each structured MIMIC-IV (Johnson et al., 2020) record into an initial textual representation by filling a predefined template. This template organizes clinically relevant attributes into a coherent note form. Specifically, the structured fields used for template construction include *Age, Gender, Ethnicity, Chief Complaint, Mode of Transport,*

Temperature, Heart Rate, Respiratory Rate, Oxygen Saturation, Blood Pressure, Pain Score, Past Medical History, and Allergies. These fields are selected to provide a relatively comprehensive characterization of the patient's condition at emergency department admission. Detailed templates are provided in Appendix A. While the resulting template-generated text t_i faithfully preserves the original clinical content, it remains stylistically rigid and lacks the fluency and expressiveness of human-written ED notes.

Retrieval. To further enhance the diversity and realism of generated triage notes, we incorporate real-world clinical documentation as stylistic guidance. Following the dataset released by TRIAGEAGENT (Lu et al., 2024), we construct a *Retrieval* component that selects relevant exemplar notes from a corpus of human-written clinical notes provided in the same benchmark. Specifically, we encode the template text and each corpus entry into a shared embedding space using a pretrained text encoder, and retrieve the top-K most similar notes based on embedding similarity.

Generation. Inspired by SDTA (Shen et al., 2025), the *Generation* component utilizes an LLM for retrieval-augmented note refinement. It combines exemplar notes with template-generated texts, allowing the model to rewrite the template-generated texts into more diverse and ideal ED notes. This process maintains the original structured clinical information while adapting to the writing style and conventions of real-world ED notes. The refined clinical triage notes m_i are then passed as input to the SOAP-Guided Encoding Module in the following subsection.

3.3 SOAP-Guided Encoding Module

Although a series of automated triage studies have been carried out, most of them primarily focus on model architectures or prompt engineering, neglecting the integration of clinical triage theory and failing to comprehensively assess patients' multidimensional clinical information. In real-world emergency triage, clinicians assess patients by reasoning through multiple clinical aspects (Bazyar et al., 2020). A widely adopted framework for this purpose is the SOAP (Subjective, Objective, Assessment, Plan) paradigm, which supports systematic clinical judgment and directly informs ESI assignment (Crausman, 1998; Jacobs, 2009; Reznich et al., 2010). Motivated by this observation, as shown in Figure 2(b), the SOAP-Guided Encoding (SGE) module is designed to explicitly incorporate the SOAP paradigm into the representation learning of patients' conditions, guiding LLMs to assess patients from four complementary perspectives and providing a more reliable basis for downstream triage inference. In addition, prior work has shown that intermediate layers of LLMs often encode richer semantic information that is more transferable to downstream tasks, whereas later layers tend to be more specialized to the pretraining objective and thus less general-purpose (Skean et al., 2025; Men et al., 2025; Liu and Niehues, 2025). Therefore, the SGE module extracts mid-depth representations from a frozen LLM, rather than using final-layer outputs, to obtain more informative SOAP-specific embeddings.

Specifically, given an ED note produced by the CNA module, the SGE module prompts a frozen LLM to analyze the case from four complementary SOAP perspectives. Subjective (S) captures patient-reported information such as chief complaints and symptoms. Objective (O) focuses on measurable and observable clinical evidence, including vital signs, examination findings, and relevant medical history. Assessment (A) reflects the clinician's integrated judgment of disease severity based on the available information. Plan (P) represents the anticipated actions and resource allocations, which play a central role in ESI determination. To analyze the ED triage note with SOAP guidance, the SGE module constructs a dialog-style prompt for each SOAP perspective. Each prompt consists of a system prompt that defines the triage task and the target SOAP perspective, followed by a user prompt containing the ED note and an instruction to analyze it from that perspective. Each prompt is processed independently by a frozen LLM. Let $C = \{S, O, A, P\}$ denote the set of SOAP perspectives, and let $c \in C$ denote a specific perspective. Given the perspective-specific prompt r_c, we extract the token-level hidden states of the i-th triage note from the l-th layer of the LLM, denoted by $\mathbf{H}_{i,c}^{(l)}$:

$$\mathbf{H}_{i,c}^{(l)} = \mathrm{LLM}_l(r_c, m_i)\,, \tag{1}$$

where $\mathrm{LLM}_l(\cdot)$ denotes the computation of the LLM up to the l-th layer, m_i refers to the i-th triage note generated by the CNA module. Full prompts are provided in Appendix G. From the model, we select a group of intermediate layers and aggregate their hidden states using a layer-mixing strategy to produce a fixed-length embedding, which is then

Dimension	MIMIC-IV				Reference				Diff
	Expert 1	Expert 2	Expert 3	Avg	Expert 1	Expert 2	Expert 3	Avg	$\Delta\text{Avg}_{\text{MI}}$
Clinical Consistency	92.19	90.94	94.69	92.60	90.00	90.00	96.00	92.00	+0.60
Factual Correctness	90.63	95.62	94.69	93.65	96.00	94.00	97.00	95.67	-2.02
Narrative Naturalness	91.25	96.88	93.12	93.75	88.00	92.00	98.00	92.67	+1.08
Information Completeness	90.63	95.00	94.06	93.23	92.00	95.00	94.00	93.67	-0.44
Readability & Clarity	91.25	93.12	94.69	93.02	91.00	90.00	98.00	93.00	+0.02
Overall	91.19	94.31	94.25	93.25	91.40	89.40	96.60	92.47	+0.78

Table 1: Human expert evaluation of constructed triage notes from MIMIC-IV, compared with real-world ED clinical notes (Reference), across five quality dimensions. Scores are linearly rescaled from a 1-5 Likert scale to a 0-100 range.

normalized. For each clinical note m_i and SOAP perspective c, we denote the resulting normalized embedding as $\mathbf{f}_{i,c} \in \mathbb{R}^d$, and feed it into downstream modules for ESI prediction.

3.4 SOAP-Aware Aggregation and Inference Module

In real-world emergency triage, clinicians often assign different importance to clinical signals based on the triage context, including information availability, symptom presentation, and perceived risk (Hall, 2011; Sanabria et al., 2020). In addition, clinicians often make a coarse risk assessment to prioritize rapid identification of critical cases before refining fine-grained ESI levels (Hinson et al., 2019). Motivated by these observations, as shown in Figure 2(c), the SAAI module is designed to first adaptively aggregate SOAP-aware representations with different weights, and then implement a two-stage hierarchical inference strategy to predict the final ESI level.

Specifically, given the four SOAP-guided representations from the SGE module, the SAAI module employs a gating network to estimate an importance weight for each component. We concatenate the four embeddings $\mathbf{f}_{i,c}$ to form $\mathbf{x}_i \in \mathbb{R}^{4d}$, based on which the gating network produces perspective-wise importance weights:

$$\mathbf{w}_i = \mathrm{softmax}\big(\mathrm{MLP}(\mathbf{x}_i)\big), \tag{2}$$

where $\mathrm{MLP}(\cdot)$ is a two-layer fully connected network with a ReLU activation and $\mathbf{w}_i \in \mathbb{R}^4$ denotes the resulting perspective-wise weight vector. These weights are then used to aggregate the SOAP embeddings into a unified representation $\mathbf{z}_i$:

$$\mathbf{z}_i = \sum_{c \in C} w_{i,c}\, \mathbf{f}_{i,c}, \tag{3}$$

where $w_{i,c}$ is the weight for perspective c, i.e., the c-th element of $\mathbf{w}_i$. Based on the aggregated representation $\mathbf{z}_i$, the SAAI module performs ESI prediction via a hierarchical inference strategy. The SAAI module first performs a coarse risk assessment that distinguishes potentially high-acuity cases from lower-acuity ones. This coarse assessment is implemented as a lightweight binary classifier and is used solely to guide subsequent inference. Conditioned on the coarse risk outcome, the SAAI module then applies a corresponding ordinal prediction head to determine the final ESI level within the predicted risk group. The high-risk and low-risk heads share the same aggregated representation while maintaining separate parameters, allowing specialization across different acuity ranges.

4 Experiments

We conduct extensive experiments to evaluate the performance of SOAPTriage by answering the following key research questions:

- **RQ1**: Does SOAPTriage outperform state-of-the-art baselines for automated ESI prediction?
- **RQ2**: Does SGE and SAAI improve the performance of SOAPTriage?
- **RQ3**: How well does SOAPTriage perform across different backbone LLM scales?
- **RQ4**: How do different architectural settings (e.g., representation depth) influence the performance of SOAPTriage?

4.1 Datasets and Preprocessing

We evaluate automated ESI prediction on a real-world emergency department dataset consisting of 15,393 ED visits, which is constructed from

MIMIC-IV as described in Section 3.2. Following CMExam and MedIQA (Liu et al., 2023; Xun et al., 2025), we split the dataset into training, validation, and test sets with a ratio of 8:1:1 while preserving the original label distribution. We also conduct a human evaluation, in which three medical-domain experts independently assess sampled across five dimensions: Clinical Consistency, Factual Correctness, Narrative Naturalness, Information Completeness, and Readability & Clarity. As shown in Table 1, the results indicate that the generated data exhibit strong clinical quality and plausibility. Additional qualitative case examples and more detailed expert evaluations of the dataset are provided in Appendix A.

4.2 Evaluation Metrics

In light of our approach's objective of automated ESI prediction, which requires both overall accuracy and clinically safe decision-making, we assess the performance of all the approaches from three distinct aspects: (1) overall prediction accuracy, measuring whether the predicted ESI level matches the ground-truth label; (2) directional triage errors, characterizing the tendency to under-triage or over-triage patients across different acuity levels; and (3) clinically significant triage errors, focusing on particularly severe under- or over-triage cases that may lead to high-stakes unsafe or inefficient clinical outcomes.

Following the evaluation protocol adopted in prior work (Lu et al., 2024), we use *Total Discordance* as the primary metric, which measures the proportion of visits for which the predicted ESI level differs from the ground-truth label, with lower values indicating better performance. To further characterize clinically relevant error patterns, we additionally report four auxiliary metrics: *UnderTriage* and *OverTriage*, which capture the frequency of severity underestimation and overestimation, respectively, as well as *Significant UnderTriage* and *Significant OverTriage*, which focus on more severe misclassifications associated with potentially unsafe triage decisions. Detailed definitions of all evaluation metrics, including the meaning of significant in the clinically motivated error metrics, are provided in Appendix C.

4.3 Baselines and Implementation Details

Baselines. We compare our method with classic methods as well as state-of-the-art models, including the following categories. *Prompting-based LLM baselines* include standard prompting (Xu et al., 2025), retrieval-augmented generation (Prompt-RAG) (Lewis et al., 2020), chain-of-thought (CoT) prompting (Kojima et al., 2022), self-consistency (SCons) (Wang et al., 2023), self-contrast (SCtr) (Wang et al., 2024), exchange-of-thought (EoT) (Yin et al., 2023), knowledge-evolvable assistant (KEA) (Lu et al., 2025) and task adaptation and instruction tuning (TAIT-LoRA) (Shen et al., 2025). *Multi-agent methods* are represented by TRIAGEAGENT (Lu et al., 2024), which integrates multiple reasoning perspectives and external evidence for triage decision-making. *Encoder-based classification models* include BERT (Devlin et al., 2019) as well as domain-specific variants such as TCM-BERT (Yao et al., 2019), BioBERT (Lee et al., 2020) and KATE-BERT (Ivanov et al., 2021), which are fine-tuned for ESI level prediction from medical notes. Detailed descriptions of baselines are provided in Appendix E.1.

Implementation Details. We implement SOAP-Triage using *Qwen3-8B* as the backbone large language model (LLM), which consists of 36 transformer layers. In the SGE module, we extract hidden representations from intermediate layers 18-22 and apply L2 normalization to obtain stream-level embeddings. During retrieval in the CNA module, we use *bge-large-en-v1.5* as the encoder model and set the retrieval top-K to 3. All models are trained with the AdamW optimizer, using a weight decay of 5×10^{-4} and an initial learning rate of 3×10^{-4}, together with a cosine annealing schedule. Training is performed for 200 epochs with a batch size of 128, and gradient clipping with a maximum norm of 5.0 is applied to stabilize optimization. For baseline implementation, standard prompting-based approaches (Standard Prompting, Prompt-RAG, CoT, SCons, SCtr, EoT) and TRIAGEAGENT are implemented using DeepSeek-V3.2 as a general-purpose instruction-following LLM. Methods requiring open-weight fine-tuning (KEA, TAIT-LoRA) and our proposed SOAPTriage are implemented using Qwen3-8B to ensure reproducibility. All experiments are conducted on a single NVIDIA RTX 5090 GPU, with the random seed fixed to 42 for reproducibility. Due to space limitations, we provide experimental details and parameters in Appendix E.2. All source code, datasets, LLM-generated embeddings, and SOAPTriage model checkpoints are publicly available in a GitHub repository.

Method	Total Discordance↓	UnderTriage↓	Significant UnderTriage↓	OverTriage↓	Significant OverTriage↓
Prompting-based LLM baselines					
Prompt	66.36	44.24	21.13	22.12	0.39
Prompt-RAG	60.56	40.41	22.94	20.14	0.33
CoT (1-Agent)	48.75	29.45	19.96	19.30	0.59
SCons (5-Agent)	45.51	26.45	20.58	19.06	0.26
SCtr (1-Agent)	56.99	37.68	29.80	19.30	0.66
EoT (3-Agent)	49.17	29.21	22.19	19.96	0.53
KEA	51.85	26.70	15.33	25.15	0.52
TAIT-LoRA (8B)	41.19	22.42	13.90	18.77	<u>0.19</u>
Multi-agent Methods					
TRIAGEAGENT	55.86	35.17	14.42	20.69	0.59
Encoder-based Methods					
BERT	42.88	22.22	15.14	20.66	0.32
TCM-BERT	40.09	21.63	13.06	<u>18.45</u>	**0.13**
BioBERT	41.71	<u>18.71</u>	13.58	23.00	**0.13**
KATE-BERT	<u>39.63</u>	18.97	<u>11.89</u>	20.66	0.32
SOAPTriage (8B)	**35.99** (3.64↓)	**17.99** (0.72↓)	**10.39** (1.50↓)	**18.00** (0.45↓)	**0.13** (0.06↓)

Table 2: Performance comparison (%) between SOAPTriage and baseline methods on the clinical triage dataset. Best results are highlighted in bold, while the second-best results are indicated by underlining. Lower values indicate better performance.

4.4 Overall Comparison (RQ1)

The experimental results on automated ESI prediction are summarized in Table 2, we have several observations.

First, encoder-based methods (e.g., BERT and TCM-BERT) generally outperform prompting-based approaches such as CoT, as well as multi-agent methods (e.g., TRIAGEAGENT). This is probably because LLMs are primarily trained on general-domain data and are not explicitly adapted to domain-specific triage scenarios. In contrast, encoder-based models trained directly for classification can learn task-specific decision boundaries more effectively from triage data. As a result, relying solely on prompt-based reasoning is insufficient to ensure reliable classification decisions in diverse and complex real-world clinical settings.

Moreover, our approach consistently outperforms all compared baseline methods. Although KATE-BERT and TAIT-LoRA are trained on triage-domain data, they are not explicitly guided by clinically grounded triage theories, which limits their ability to systematically capture comprehensive and clinically critical information. In contrast, our model is not only specifically optimized for the triage domain, but also performs inference under the guidance of the SOAP clinical framework. This explicit integration of clinical reasoning enables our approach to achieve more reliable and accurate triage predictions than existing methods.

4.5 Ablation Study (RQ2)

To analyze the rationality and effectiveness of the designed components in our SOAPTriage framework, we conduct an ablation study to compare SOAPTriage with its three variants: (1) *w/o SGE*, which ignores the SOAP paradigm by removing SGE and directly encoding the note using the direct-triage prompt; (2) *w/o SAAI-weight*, which removes the weighting scheme and directly aggregates the SOAP views by assigning uniform weights to each view; (3) *w/o SAAI-fusion*, which removes multi-view fusion by directly using only a single randomly sampled SOAP view.

The results presented in Figure 3(a) demonstrate the following findings. (1) The SGE module benefits ESI prediction performance. We note a notable numerical increase in *Total Discordance* and other evaluation metrics when removing SGE module. This phenomenon underscores the critical role of the SOAP paradigm in structuring clinically meaningful representations for triage inference. (2) The weighting mechanism in the SAAI module can al-

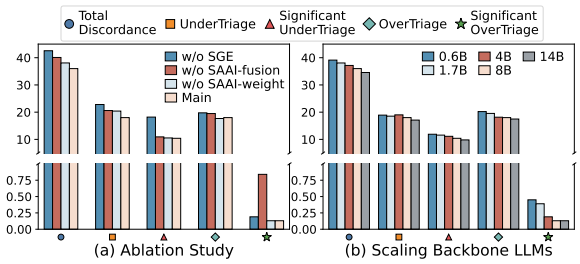

Figure 3: Ablation study and performance comparison across backbone LLMs of different scales. Lower values indicate better performance.

Model	Total ↓	Under ↓	S-Under ↓	Over ↓	S-Over ↓
Main	**35.99**	**17.99**	**10.39**	**18.00**	**0.13**
Single stage	38.08	18.91	13.12	19.17	0.45
Last layer	37.69	18.00	10.98	19.69	0.39

Table 3: Performance comparison (%) of three variants under the Qwen3-8B setting across five triage metrics. We report Total Discordance, UnderTriage, and OverTriage rates, as well as their clinically significant counterparts (*S-Under* and *S-Over*). Lower values indicate better performance.

leviate the issue of *Significant OverTriage*. Comparing the values in the fifth metric column in Figure 3(a), we observe that in the absence of SAAI weighting mechanism, the *Significant OverTriage* metric experiences an increase. (3) The multi-view inference mechanism not only substantially reduces *Total Discordance*, but also helps mitigate *UnderTriage*. It is evident that the absence of the multi-view inference leads to a significant numerical increase in *UnderTriage* and *Total Discordance*.

4.6 Effectiveness of Scaling Backbone LLMs (RQ3)

We conduct several experiments to explore the effectiveness of the proposed SOAPTriage framework across backbone LLMs of varying scales, including models with 0.6B, 1.7B, 4B, 8B, and 14B parameters. As shown in Figure 3(b), we observe that models with larger parameter scales outperform those with fewer parameters across evaluation metrics. This trend is likely because larger models provide greater representational capacity, enabling the SGE module to capture richer and more nuanced clinical semantic information. Moreover, considering both performance gains and computational cost, the 8B model offers a favorable trade-off, achieving strong performance while maintaining practical efficiency. Accordingly, we adopt it as the backbone model for SOAPTriage in subsequent experiments.

4.7 Influence of Different Architectural Settings (RQ4)

We want to explore the impact of different architectural settings on the performance of SOAPTriage, and therefore design two model variants: (1) *Single-stage inference*, which replaces the two-stage hierarchical inference with direct ESI prediction; and (2) *Last-layer representation*, which replaces the mid-depth representations with those from the final backbone layer. From the results in Table 3, we draw two observations. First, the single-stage variant performs worse overall and exhibits larger performance variance than the full coarse-to-fine setting, suggesting that explicitly modeling the hierarchical triage process in accordance with clinicians' triage practice is consistently beneficial in our setting. Second, replacing mid-depth embeddings with last-layer embeddings leads to a moderate decline in overall performance stability, suggesting that mid-depth representations may be more suitable for this task.

5 Conclusion

Automated ESI prediction is critical yet challenging, as existing approaches are hindered by limited access to high-quality triage data and the lack of a clinically grounded triage framework. This paper introduces SOAPTriage, a SOAP-guided multi-view clinical text modeling framework for automated ESI prediction that integrates a large-scale dataset with a SOAP-guided modeling paradigm. By constructing a natural-language triage dataset and explicitly modeling the Subjective, Objective, Assessment, and Plan perspectives, SOAPTriage captures complementary clinical views and integrates them through adaptive aggregation and hierarchical inference. Extensive empirical analyses demonstrate the effectiveness of SOAPTriage in improving ESI prediction performance. A key takeaway from this work is that explicitly incorporating clinically established reasoning paradigms, such as SOAP, is essential for reliable automated triage. Furthermore, extending SOAP-guided reasoning to more complex clinical scenarios, such as multimodal inputs or longitudinal patient records, could enable richer representations and provide more comprehensive clinical decision-making support in future work.

Limitations

Despite the encouraging results achieved by SOAPTriage, several limitations remain. First, all samples used and generated in this work are in English. Extending SOAPTriage to other languages would require additional language-specific preprocessing. Second, due to computational and cost constraints, our experiments are limited to backbone models ranging from 0.6B to 14B parameters. Exploring larger-scale models, such as those with 30B-70B parameters, is an important direction for future work. Third, although SOAPTriage attempts to model the triage reasoning process by incorporating clinically grounded triage theories, the underlying task formulation remains a classification problem, which inherently limits interpretability. While recent studies have begun to decompose ESI prediction into explicit decision points to enhance transparency (Shen et al., 2025), such approaches typically rely on extensive expert annotations, making large-scale and automated evaluation challenging. In future work, we plan to investigate self-consistent reasoning supervision strategies to enable scalable assessment of decision-level reasoning without incurring prohibitive annotation costs. Finally, although our dataset is constructed to broadly cover diverse patients and ESI levels, inherent biases and limitations in both the data and the models cannot be fully eliminated. As a result, the system may not fully capture all forms of patient diversity or address every clinical context equally well.

Ethical Considerations

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. All datasets and derived artifacts in this work are used solely for research purposes. This work aims to advance automated Emergency Severity Index (ESI) prediction by integrating clinically grounded triage theories with LLMs, with the goal of supporting, rather than replacing, clinical decision-making in emergency department triage. We recognize several ethical considerations associated with this line of research. First, although the proposed framework leverages large-scale synthetic triage notes constructed from public structured records, the generated data may still reflect biases present in the original sources and in the language models used for generation. Second, automated triage systems, including SOAPTriage, should not be deployed as standalone decision-makers in clinical practice. Model outputs are intended solely as decision support and require appropriate clinical oversight. Finally, while our experiments demonstrate improved performance and stability, automated triage remains a high-stakes application. Careful validation, continuous monitoring, and alignment with clinical governance are necessary before any real-world deployment. We hope this work encourages further research on responsible integration of clinical reasoning frameworks and machine learning models in safety-critical healthcare settings.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62376130, No.62576120), Taishan Scholars Program (No.TSPD20240814), Program of New Twenty Policies for Universities of Jinan (No.202333008), the Open Project of the Key Laboratory of Computing Power Network and Information Security, Ministry of Education (No.2024ZD020), the Pilot Project for Integrated Innovation of Science, Education, Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01), and Shandong Talent Introduction Program (No.WSR2025005).

References

Jafar Bazyar, Mehrdad Farrokhi, Amir Salari, and Hamid Reza Khankeh. 2020. The principles of triage in emergencies and disasters: A systematic review. *Prehospital and Disaster Medicine*, 35(3):305–313.

Robet S Crausman. 1998. The ethics SOAP note. *Chest*, 113(2):1–1.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2025. AugGPT: Leveraging ChatGPT for text data augmentation. *IEEE Transactions on Big Data*, 11(3):907–918.

Adebayo Da'Costa, Jennifer Teke, Joseph E Origbo, Ayokunle Osonuga, Eghosasere Egbon, and David B Olawade. 2025. AI-driven triage in emergency departments: A review of benefits, challenges, and future directions. *International Journal of Medical Informatics*, 197(1):1–7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4171–4186.

Rabie Adel El Arab and Omayma Abdulaziz Al Moosa. 2025. The role of AI in emergency department triage: An integrative systematic review. *Intensive and Critical Care Nursing*, 89(1):1–10.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ari B Friedman, M Kit Delgado, and Gary E Weissman. 2024. Artificial intelligence for emergency care Triage-Much promise, but still much to learn. *JAMA Network Open*, 7(5):e248857–e248857.

Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. 2022. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *arXiv preprint arXiv:2211.10330*, pages 1–21.

Judith A Hall. 2011. Clinicians' accuracy in perceiving patients: Its relevance for clinical practice and a narrative review of methods and correlates. *Patient Education and Counseling*, 84(3):319–324.

Jeremiah S Hinson, Diego A Martinez, Stephanie Cabral, Kevin George, Madeleine Whalen, Bhakti Hansoti, and Scott Levin. 2019. Triage performance in emergency medicine: A systematic review. *Annals of Emergency Medicine*, 74(1):140–152.

Oleksandr Ivanov, Lisa Wolf, Deena Brecher, Erica Lewis, Kevin Masek, Kyla Montgomery, Yurii Andrieiev, Moss McLaughlin, Stephen Liu, Robert Dunne, and 1 others. 2021. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. *Journal of Emergency Nursing*, 47(2):265–278.

Lee Jacobs. 2009. Interview with Lawrence Weed, MD-the father of the problem-oriented medical record looks ahead. *The Permanente Journal*, 13(3):1–6.

Sarah Jaroudi and J Drew Payne. 2019. Remembering Lawrence Weed: A pioneer of the SOAP note. *Academic Medicine*, 94(1):1–1.

A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. 2023a. MIMIC-IV-ED (version 2.2).

A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. 2023b. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2).

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. MIMIC-IV. *PhysioNet*, pages 49–55.

Joshua W Joseph, Evan L Leventhal, Anne V Grossestreuer, Matthew L Wong, Loren J Joseph, Larry A Nathanson, Michael W Donnino, Noémie Elhadad, and Leon D Sanchez. 2020. Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. *Journal of the American College of Emergency Physicians Open*, 1(5):773–781.

Eleni Karlafti, Athanasios Anagnostis, Theodora Simou, Angeliki Sevasti Kollatou, Daniel Paramythiotis, Georgia Kaiafa, Triantafyllos Didaggelos, Christos Savvopoulos, and Varvara Fyntanidou. 2023. Support systems of clinical decisions in the triage of the emergency department using artificial intelligence: the efficiency to support triage. *Acta Medica Lituanica*, 30(1):1–7.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Saikam Law, Brian Oldfield, Wah Yang, and Global Obesity Collaborative. 2024. ChatGPT/GPT-4 (large language models): Opportunities and challenges of perspective in bariatric healthcare professionals. *Obesity Reviews*, 25(7):1–6.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael Mahoney, Kurt Keutzer, and Amir Gholami. 2024. LLM2LLM: Boosting LLMs with novel iterative data enhancement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 6498–6526.

David M Levine, Rudraksh Tuwani, Benjamin Kompa, Amita Varma, Samuel G Finlayson, Ateev Mehrotra, and Andrew Beam. 2024. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: An observational study. *The Lancet Digital Health*, 6(8):e555–e561.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 9459–9474.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(33):5–55.

Michelle P. Lin, Dhruv Sharma, Arjun Venkatesh, Stephen K. Epstein, Alexander Janke, Nicholas Genes, Abhi Mehrotra, James Augustine, Bill Malcolm, Pawan Goyal, and Richard T. Griffey. 2024. The clinical emergency data registry: Structure, use, and limitations for research. *Annals of Emergency Medicine*, 83(5):467–476.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. *arXiv preprint arXiv:2502.14830*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on CMExam-a comprehensive Chinese medical exam dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 52430–52452.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing*, pages 5747–5764.

Wenpeng Lu, Kangjun Liu, Jianlei Wang, Xueping Peng, Tao Shen, Fa Zhu, Weiyu Zhang, Jiabing Zhu, Tao Xin, and Athanasios V. Vasilakos. 2025. Advancing Chinese conversation-based patient guidance with a benchmark and knowledge-evolvable assistant. *IEEE Journal of Biomedical and Health Informatics*, 1(1):1–12.

Guozheng Ma, Zhen Wang, Zhecheng Yuan, Xueqian Wang, Bo Yuan, and Dacheng Tao. 2025. A comprehensive survey of data augmentation in visual reinforcement learning. *International Journal of Computer Vision*, 133(10):7368–7405.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. ShortGPT: Layers in large language models are more redundant than you expect. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 20192–20204.

Sophie Rentschler, Martin Riedl, Christian Stab, and Martin Rückert. 2022. Data augmentation for intent classification of German conversational agents in the finance domain. In *Proceedings of the 18th Conference on Natural Language Processing*, pages 1–7.

Christopher B Reznich, Dianne P Wagner, and Mary M Noel. 2010. A repurposed tool: The programme evaluation SOAP note. *Medical Education*, 44(3):298–305.

Andrea Juliana Sanabria, Hector Pardo-Hernandez, Mónica Ballesteros, Carlos Canelo-Aybar, Emma McFarlane, Ena Niño de Guzman, Katrina Penman, Margarita Posso, Marta Roqué i Figuls, Anna Selva, Robin W. M. Vernooij, Pablo Alonso-Coello, Laura Martínez García, Arnav Agarwal, Sophie Blanchard, Laura Brereton, Melissa Brouwers, Itziar Etxeandia-Ikobaltzeta, Ivan D. Flórez, and 13 others. 2020. The UpPriority tool was developed to guide the prioritization of clinical guideline questions for updating. *Journal of Clinical Epidemiology*, 126(1):80–92.

Rocío Sánchez-Salmerón, José L Gómez-Urquiza, Luis Albendín-García, María Correa-Rodríguez, María Begoña Martos-Cabrera, Almudena Velando-Soriano, and Nora Suleiman-Martos. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60(1):1–7.

Martin Scherer, Dagmar Lühmann, Agata Kazek, Heike Hansen, and Ingmar Schäfer. 2017. Patients attending emergency departments: A cross-sectional study of subjectively perceived treatment urgency and motivation for attending. *Deutsches Ärzteblatt International*, 114(39):1–9.

Qingyang Shen, Xiaozhi Zhang, Haomin Ren, Quan Guo, and Zhang Yi. 2025. Knowledge-embedded large language models for emergency triage. *Knowledge-Based Systems*, 318(1):1–13.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 1–22.

Chenxu Wang, Fei Wang, Shuhan Li, Qing-wen Ren, Xiaomei Tan, Yaoyu Fu, Di Liu, Guangwu Qian, Yu Cao, Rong Yin, and 1 others. 2025. Patient triage and guidance in emergency departments using large language models: Multimetric study. *Journal of Medical Internet Research*, 27(1):1–16.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency improves chain of thought reasoning in language models. In *Proceedings of the 19th International Conference on Learning Representations*, pages 1–24.

Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. 2026. A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 38(1):47–66.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics, pages 257–279.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing*, pages 6382–6388.

Christopher YK Williams, Brenda Y Miao, Aaron E Kornblith, and Atul J Butte. 2024a. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications*, 15(1):1–10.

Christopher YK Williams, Travis Zack, Brenda Y Miao, Madhumita Sushil, Michelle Wang, Aaron E Kornblith, and Atul J Butte. 2024b. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open*, 7(5):e248895–e248895.

Richard C Wuerz, Debbie Travers, Nicki Gilboy, David R Eitel, Alex Rosenau, and Ramine Yazhari. 2001. Implementation and refinement of the emergency severity index. *Academic Emergency Medicine*, 8(2):170–176.

Lei Xu, Wenzhe Zhao, and Xin Huang. 2025. Diagnosis and triage performance of contemporary large language models on short clinical vignettes. *Journal of Medical Systems*, 49(1):1–12.

Siyi Xun, Yue Sun, Jingkun Chen, Zitong Yu, Tong Tong, Xiaohong Liu, Mingxiang Wu, and Tao Tan. 2025. MedIQA: A scalable foundation model for prompt-driven medical image quality assessment. In *Proceedings of the 28th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. IM-RAG: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 730–740.

Li-Hung Yao, Ka-Chun Leung, Chu-Lin Tsai, Chien-Hua Huang, and Li-Chen Fu. 2021. A novel deep learning–based system for triage in the emergency department using electronic medical records: retrospective cohort study. *Journal of Medical Internet Research*, 23(12):1–15.

Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang, and Yuan Luo. 2019. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *Journal of the American Medical Informatics Association*, 26(12):1632–1636.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153.

Zaifu Zhan, Jun Wang, Shuang Zhou, Jiawen Deng, and Rui Zhang. 2025. MMRAG: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning. *Journal of the American Medical Informatics Association*, 32(10):1505–1516.

Kaiyuan Zhang, Qian Liu, Luyang Zhang, Chaoqun Zheng, Shuaimin Li, Bing Xu, Muyun Yang, Xinxiao Qiao, and Wenpeng Lu. 2025a. MADAWSD: Multi-Agent debate framework for adversarial word sense disambiguation. In *Proceedings of the 30th Empirical Methods in Natural Language Processing*, pages 22294–22313.

Luyang Zhang, Shuaimin Li, Yishuo Li, Kunpeng Kang, Kaiyuan Zhang, Cong Wang, and Wenpeng Lu. 2025b. RoDEval: A robust word sense disambiguation evaluation framework for large language models. In *Proceedings of the 30th Empirical Methods in Natural Language Processing*, pages 17095–17126.

A Dataset Construction Details

To evaluate the generalizability of SOAPTriage beyond a single data source, the raw clinical records used for data augmentation are drawn from two large-scale, publicly available and de-identified datasets.

- **MIMIC-IV**. We primarily conduct our experiments on MIMIC-IV. We extract eligible emergency department (ED) visits by integrating MIMIC-IV (Johnson et al., 2020), MIMIC-IV-ED (Johnson et al., 2023a), and MIMIC-IV-Note (Johnson et al., 2023b), which provide complementary information to construct a comprehensive triage dataset.
- **NHAMCS**. We additionally use the National Hospital Ambulatory Medical Care Survey (NHAMCS)[2] as an external source of ED visit records covering a broad range of triage scenarios, with additional experiments reported in Appendix F.1.

Inspired by TRIAGEAGENT (Lu et al., 2024), we identify and retain a subset of triage-relevant metadata, including patient age, visit date, sex, and chief complaint, which serve as the basis for subsequent data construction. We next present the triage composition of our dataset and describe the predefined template used in the Templating stage of the Clinical Note Augmentation module, and provide representative examples illustrating the Retrieval and Generation stages. In addition, in Section A.4, we report human expert evaluation results to assess the clinical fidelity of the triage notes generated from both datasets. Since NHAMCS adopts the IMMEDR (hereafter referred to as the immediacy rating, IR) triage rating, which is comparable to the ESI scale, we further evaluate the agreement between IR and ESI.

A.1 Data Statistics

The augmented MIMIC-IV dataset contains 15,393 records, while the NHAMCS dataset contains 16,596 records. We split each dataset into training, validation, and test sets using an 8:1:1 ratio. Detailed dataset statistics are reported in Table 4.

Dataset	ESI-1	ESI-2	ESI-3	ESI-4	ESI-5
			MIMIC-IV		
Train	1170	3656	5850	1280	359
Validation	146	457	731	160	45
Test	146	457	731	160	45
Dataset	**IR-1**	**IR-2**	**IR-3**	**IR-4**	**IR-5**
			NHAMCS		
Train	378	2400	5760	3861	887
Validation	45	300	720	480	110
Test	45	300	720	480	110

Table 4: Dataset statistics of the constructed triage datasets.

A.2 Template

Predefined Template. To convert structured ED records into an initial textual representation, we design a unified template that verbalizes key triage-relevant fields into a coherent clinical narrative. The predefined template is illustrated in Figure 4.

Predefined Template

This is a clinical case involving a {age_val}-year-old {gender_val} of {race_val} ethnicity, who presented to the emergency department. The patient's primary concern was: "{cc}".
The patient arrived via {transport}. Initial vital signs were: temperature {temp}°F, heart rate {hr}, respiratory rate {rr}, SpO_2 {o2}%, blood pressure {sbp}/{dbp} mmHg. Pain was assessed at {pain}/10.
Relevant medical history includes: {pmh}. The patient reported the following allergies: {allergies}.

Figure 4: Predefined template used in the Clinical Note Augmentation module.

Prompt. As shown in Figure 5, we design a prompt that leverages retrieved real-world clinical notes to refine and polish the template-filled text.

A.3 Illustrative Examples

As shown in Figure 6, we present a representative example of the Clinical Note Augmentation process. The figure illustrates how retrieved real-world ED clinical notes are used as in-context exemplars to guide an LLM in refining a template-filled ED text into a polished natural-language triage note that preserves the original clinical content while

[2] https://www.cdc.gov/nchs/nhamcs/about/index.html

Prompt Design for the Clinical Note Augmentation Module

System Prompt:
Please polish and improve the following emergency department patient description to make it more natural, fluent, and clinically appropriate, similar to the writing style of the examples provided.

Examples:
1. "{match_text1}"
2. "{match_text2}"
3. "{match_text3}"

Original Text To Polish:
1. "{template_txt}"

Task:
Please rewrite and expand the original text using a more natural and professional medical note style while retaining all key clinical information. Output only the polished version without any additional explanations.

Figure 5: Prompt design for the Clinical Note Augmentation module in the SOAPTriage framework.

adopting realistic clinical writing style. This process reduces the rigidity of template-based text and encourages more diverse, human-like clinical narratives.

A.4 Human Expert Evaluation

To rigorously assess the quality, clinical plausibility, and faithfulness of the constructed datasets, we conducted a multi-faceted human expert evaluation involving three medical-domain experts. The evaluation focuses on three primary aspects: overall clinical quality, hallucination-related error detection, and IMMEDR-ESI mapping alignment. Annotating each sample required approximately 15-30 minutes, with an average annotation time of 22 minutes. Annotators were compensated at an hourly rate of $3.15, in compliance with local labor regulations. The corresponding annotation interfaces are illustrated in Figures 7, 8, and 9.

Clinical Quality and Plausibility. We evaluated the constructed triage notes by randomly sampling 80 instances from each of our two datasets, along with 20 notes from a reference set of real-world emergency department (ED) clinical notes. Three medical-domain experts independently assessed each note across five dimensions: Clinical Consistency, Factual Correctness, Narrative Naturalness, Information Completeness, and Readability & Clarity.

Each dimension was rated on a 5-point Likert scale (Likert, 1932), which was subsequently rescaled to a 0-100 range for ease of comparison. The aggregated results, reported in Table 5, indicate that although a marginal performance gap exists relative to real-world notes, the differences remain small across all dimensions. Overall, these findings suggest that the generated triage notes exhibit a high degree of clinical plausibility and are suitable for supporting automated triage modeling.

Hallucination Assessment. To detect potential hallucination-related errors introduced during the LLM-based augmentation process, we randomly sampled 100 instances from each of the augmented MIMIC-IV and NHAMCS datasets for expert review. Experts assigned a field-level faithfulness score (1-10) based on three criteria: (i) preservation of structured fields without value alterations, (ii) absence of unsupported or fabricated information, and (iii) coverage of relevant fields without critical omissions. The results, summarized in Table 6 and rescaled to a 0-100 range for consistency, indicate that triage notes generated via the CNA module remain highly faithful to the source data. Notably, while experts observed that the model occasionally omits the *Race* field, other clinically relevant fields were consistently incorporated without apparent errors or inconsistencies.

IMMEDR-ESI Mapping Evaluation. The IMMEDR triage rating used in NHAMCS and the Emergency Severity Index (ESI) adopt different assessment strategies and operational criteria. However, both triage schemes are designed to stratify patients according to clinical severity and urgency, suggesting an inherent conceptual similarity between the two scales. Given the limited availability of large-scale, high-quality ESI-annotated datasets, we incorporate NHAMCS as a complementary data source and explore an idealized mapping from IMMEDR to ESI to enable its use for automated triage modeling. To validate the feasibility of this mapping, we conduct a human expert evaluation in which two medical-domain experts independently assign ESI triage levels to a pro-

Illustration of the Clinical Note Augmentation Process

Input:
This is a clinical case involving a 65-year-old male of white ethnicity, who presented to the emergency department. The patient's primary concern was: "shaking chills". The patient arrived via walk in. Initial vital signs were: temperature 99.2°F, heart rate 88.0, respiratory rate 18.0, SpO_2 100.0%, blood pressure 173.0/140.0 mmHg. Pain was assessed at 3/10. Relevant medical history includes: DM type 1 Gastroparesis ESRD s/p LRRT (baseline Cr 0.8-1.0) HTN Hypercholesterolemia CAD s/p NSTEMI GERD. The patient reported the following allergies: No Known Allergies / Adverse Drug Reactions.

Match_text1:
36-year-old male presents to the emergency department reporting severe abdominal pain for the past 24 hours. The patient has no Significant medical history, and the pain started in the lower right quadrant of the abdomen, progressively worsening. He rates the pain as 8/10, and reports nausea but no vomiting or fever. His vital signs are stable, with a blood pressure of 120/80 mmHg, a heart rate of 85 beats per minute, a respiratory rate of 16 breaths per minute, and an oxygen saturation of 98%. The patient denies any recent trauma or unusual physical activity. No other associated symptoms like chest pain or dizziness are reported.

Match_text2:
A 34-year-old male presents to triage with right lower quadrant pain, 5/10, all day. Pain is associated with loss of appetite, nausea and vomiting. Past medical history: None. The patient appears in moderate discomfort, skin warm and dry, guarding abdomen.

Match_text3:
A 17-year-old female presents to the ED with complaints of severe abdominal pain that started a few hours ago. She is alert and in moderate discomfort but able to answer questions. Vital signs: BP 120/80, HR 88, RR 18, T 98.7°F. She denies any recent trauma or medical history.

Output: A 65-year-old male presented to the emergency department via walk in, reporting a primary complaint of "shaking chills." He was alert and described the chills as significant, rating his overall discomfort as 3 out of 10. His initial vital signs at triage were: temperature 99.2°F, heart rate 88 beats per minute, respiratory rate 18 breaths per minute, blood pressure 173/140 mmHg, and oxygen saturation of 100% on room air. His past medical history is significant for Type 1 Diabetes Mellitus, gastroparesis, End-Stage Renal Disease status post living-related renal transplant with a baseline creatinine of 0.8 to 1.0, hypertension, hypercholesterolemia, coronary artery disease status post NSTEMI, and GERD. He reports no known drug allergies.

Figure 6: An example of clinical note refinement in SOAPTriage, illustrating the structured ED input, retrieved in-context clinical notes, and the final refined triage note. The retrieved exemplars introduce authentic stylistic and phrasal variations from real-world clinical notes, reducing the rigidity of template-based text and yielding more diverse, human-like clinical narratives.

portionally sampled set of 110 NHAMCS cases based solely on ESI guidelines. We then quantify the agreement between the expert-assigned ESI labels and the original IMMEDR ratings using Fleiss' kappa (κ) (Fleiss, 1971; Landis and Koch, 1977), which measures inter-rater consistency beyond chance. The resulting Fleiss' kappa κ of 0.7435 indicates substantial agreement, suggesting that the IMMEDR-to-ESI mapping, although it cannot fully replace direct ESI annotations, serves as a viable idealized mapping for complementary analysis.

B Emergency Severity Index (ESI)

The Emergency Severity Index (ESI) is a widely adopted five-level triage system that categorizes emergency department visits according to clinical urgency (Wuerz et al., 2001), ranging from level 1 (most urgent) to level 5 (least urgent). In practice, clinicians refer to the ESI handbook and apply their medical judgment to assess a patient's condition and assign an appropriate ESI level during triage.

- **ESI-1**: Most urgent and requires immediate life-saving intervention.
- **ESI-2**: High urgency and potentially life-threatening, requiring prompt attention.
- **ESI-3**: Urgent but not immediately life-threatening, requiring multiple diagnostic or therapeutic resources.
- **ESI-4**: Less urgent and requires a single resource, posing no immediate threat to life.
- **ESI-5**: Least urgent and requires no immediate resources, allowing for longer waiting times.

C Evaluation Metrics and Definitions

Following the evaluation protocol adopted in prior work (Lu et al., 2024), we evaluate all methods using a set of clinically motivated metrics that capture both overall accuracy and clinically relevant error patterns in emergency triage. In these metrics, significant denotes the clinical severity of triage deviations rather than statistical significance.

The *Total Discordance* rate measures the overall error rate of a model and is defined as the proportion of visits for which the predicted ESI label differs from the ground-truth label. It is computed as follows:

$$\textit{Total Discordance} = \frac{\text{Total Misclassifications}}{\text{Total number of visits}}, \quad (4)$$

where *Total misclassifications* denotes the number of visits incorrectly classified by the model, and *Total number of visits* denotes the total number of evaluated visits.

The *UnderTriage* rate is defined as the proportion of visits for which the predicted ESI level indicates lower urgency than the ground-truth label, and is computed as follows:

$$\textit{UnderTriage} = \frac{\text{Number of UnderTriage predictions}}{\text{Total number of visits}}. \quad (5)$$

Similarly, the *OverTriage* rate is defined as the proportion of visits for which the predicted ESI level indicates higher urgency than the ground-truth label, and is computed as follows:

$$\textit{OverTriage} = \frac{\text{Number of OverTriage predictions}}{\text{Total number of visits}}. \quad (6)$$

The *Significant UnderTriage* rate captures cases in which clinically critical visits with ground-truth ESI levels of 1 or 2 are incorrectly assigned lower-urgency predictions, specifically ESI levels 3, 4, or 5. It is computed as follows:

$$\textit{Significant UnderTriage} = \frac{\text{Predicted-3,4,or 5}}{\text{Total number of visits}}. \quad (7)$$

The *Significant OverTriage* rate captures cases in which visits with lower-acuity ground-truth labels (ESI 3-5) are incorrectly overestimated by the model and predicted as the most urgent level (ESI 1). It is computed as follows:

$$\textit{Significant OverTriage} = \frac{\text{Predicted-1}}{\text{Total number of visits}}. \quad (8)$$

D Case Study

To provide deeper insight into the decision behavior of SOAPTriage, we conduct a detailed case study analysis of representative triage examples. As shown in Table 7, we present three cases that span different clinical presentations and acuity levels, including both correct predictions (Case I and Case II) and typical failure scenarios (Case III). These cases illustrate how different SOAP views contribute to model predictions across triage stages, and we further examine typical failure cases to highlight the limitations and error patterns of the proposed framework.

	MIMIC-IV				NHAMCS				Reference				Diff	
Dimension	**Expert 1**	**Expert 2**	**Expert 3**	**Avg**	**Expert 1**	**Expert 2**	**Expert 3**	**Avg**	**Expert 1**	**Expert 2**	**Expert 3**	**Avg**	**ΔAvg$_{MI}$**	**ΔAvg$_{NH}$**
Clinical Consistency	92.19	90.94	94.69	92.60	82.75	90.50	97.00	90.08	90.00	90.00	96.00	92.00	+0.60	-1.92
Factual Correctness	90.63	95.62	94.69	93.65	85.50	91.50	95.75	90.92	96.00	94.00	97.00	95.67	-2.02	-4.75
Narrative Naturalness	91.25	96.88	93.12	93.75	88.25	88.75	97.00	91.33	88.00	92.00	98.00	92.67	+1.08	-1.34
Information Completeness	90.63	95.00	94.06	93.23	88.25	90.75	97.50	92.17	92.00	95.00	94.00	93.67	-0.44	-1.50
Readability & Clarity	91.25	93.12	94.69	93.02	82.00	90.50	97.25	89.92	91.00	90.00	98.00	93.00	+0.02	-3.08
Overall	91.19	94.31	94.25	93.25	85.35	90.40	96.90	90.88	91.40	89.40	96.60	92.47	+0.78	-1.59

Table 5: Human expert evaluation of constructed triage notes from MIMIC-IV and NHAMCS, compared with real-world ED clinical notes (Reference), across five quality dimensions. Scores are linearly rescaled from a 1-5 Likert scale to a 0-100 range.

Expert	**MIMIC-IV**	**NHAMCS**
Expert 1	92.80	90.30
Expert 2	95.40	91.40
Expert 3	93.80	92.00
Avg	94.00	91.23

Table 6: Human expert assessment of field-level faithfulness for triage notes generated by the CNA module on MIMIC-IV and NHAMCS. Scores are rescaled from a 1-10 scale to a 0-100 range.

D.1 Case I

In this example, both the ground-truth triage decision and the model prediction assign the visit a low acuity level (ESI-4).

Coarse-grained triage stage. At the coarse risk-routing stage, the model assigns most of the weight to Objective (46.38%) and Subjective (45.46%) views, with substantially lower contributions from Assessment (4.30%) and Plan (3.87%). This weighting pattern is consistent with the decision goal at this stage, which is to distinguish potentially high-risk cases from clearly low-risk ones rather than to determine an exact ESI level. The clinical note is primarily composed of subjective symptom descriptions, such as redness in the left eye and a pain level of 5/10. These signals directly inform whether the presentation is concerning and are therefore captured by the Subjective view. In parallel, the Objective view contributes more because stable vital signs not only confirm the absence of acute instability but also provide strong evidence for differentiating risk at this coarse stage. The Plan view encodes the expected level of diagnostic and therapeutic resources implied by the presentation, which remains low in this case, making it a supportive cue for routing the visit to the low-risk branch. Similarly, the Assessment view, which summarizes a more fine-grained clinical judgment, is not yet central to the routing decision and becomes more relevant only after the case has been assigned to a low-risk pathway.

Fine-grained triage stage. After routing to the low-risk branch, the importance distribution shifts markedly. Assessment becomes the primary contributor (57.59%), while Plan remains influential (39.30%), and both Subjective (1.44%) and Objective (1.66%) receive negligible weight. This shift aligns with the clinical logic of low-acuity triage. Once immediate risk has been excluded, the task transitions from risk screening to fine-grained ESI determination. At this stage, decision-making relies mainly on a synthesized clinical judgment of overall severity, which is captured by the Assessment view, together with anticipated resource utilization encoded in Plan. Objective findings, such as stable vital signs, primarily confirm clinical stability and offer limited discriminative value across low-acuity levels. Likewise, raw symptom descriptions in Subjective contribute little additional information, as their implications have already been abstracted into higher-level severity estimates. Consequently, the final ESI prediction in the low-risk branch is driven chiefly by the model's integrated assessment of severity and expected resource demand.

D.2 Case II

In this example, both the ground-truth triage decision and the model prediction assign the visit the highest acuity level (ESI-1).

Coarse-grained triage stage. At the coarse risk-routing stage, the model places nearly all its weight on the Subjective (52.12%) and Plan (37.11%) views, with negligible contributions from Objective (10.36%) and Assessment (0.41%). This weighting pattern aligns with the objective of this stage, which is to rapidly identify immediately

Case I: A 27-year-old female of Chinese ethnicity presents to the Emergency Department via walk in with a primary complaint of left eye redness. She is alert and cooperative at triage. Her initial vital signs are recorded as: temperature 98.2°F, heart rate 78 beats per minute, respiratory rate 18 breaths per minute, SpO_2 99% on room air, and blood pressure 106/60 mmHg. She reports her pain level as 5 out of 10.
Case II: A 68-year-old male arrived by ambulance to the emergency department. At triage, his primary concern was reported as "iph, transfer." His vital signs were recorded as HR 77, BP 141/78. He reports a past medical history of diabetes mellitus and a prior TIA. He states his allergies include ACE inhibitors, penicillins, quinine, and sildenafil.
Case III: An 81-year-old white female presented to the emergency department via walk in. At triage, she stated her primary concern was nausea and vomiting. Her initial vital signs were recorded as: temperature 98.8°F, heart rate 110 beats per minute, respiratory rate 22 breaths per minute, blood pressure 120/95 mmHg, and oxygen saturation 94% on room air. She reported a pain level of 0/10. Her past medical history is significant for COPD, a history of hepatic abscess, severe mitral regurgitation with an ejection fraction of 45%, a history of DVT, restless leg syndrome, status post cholecystectomy, and a history of melanoma. She has undergone frequent ERCPs for large common bile duct stones, with the most recent procedure involving metallic stent placement for a biliary stricture. She reports a latex allergy.

Table 7: Three representative emergency department triage cases used for qualitative analysis.

life-threatening cases. In this case, the Subjective view is highly informative because the chief complaint ("IPH, transfer") immediately signals a time-sensitive neurologic emergency. Meanwhile, the Plan view captures the anticipated need for urgent, resource-intensive actions—rapid neurologic evaluation, emergent neuroimaging, laboratory testing (including coagulation studies), continuous monitoring, and potential blood pressure management—thereby reinforcing the high-risk routing decision. By contrast, the Objective view contributes less at this stage because the available vital signs (HR 77, BP 141/78) do not by themselves indicate immediate physiologic collapse.

Fine-grained triage stage. After routing to the high-risk branch, the importance distribution shifts. Assessment becomes the largest contributor (41.34%), while Plan (29.00%) and Objective (16.40%) remain informative and Subjective (13.26%) provides limited supplementary evidence. This transition reflects the change in decision focus from risk identification to precise acuity confirmation within the high-risk category. Once the case is established as critical, the final ESI assignment relies more heavily on an integrated clinical judgment of severity, which is encoded in the Assessment view. Objective signals continue to support this decision but no longer need to dominate, as their primary role in triggering high-risk routing has already been fulfilled. Similarly, the contribution of the Plan view decreases because, once a case is identified as high risk, substantial resource utilization is implicitly expected and therefore provides limited additional discrimination among critical acuity levels. As a result, the final ESI-1 prediction is driven mainly by the model's synthesized assessment of overall clinical severity, supported by objective evidence and anticipated care requirements.

D.3 Case III

In this example, the ground-truth triage label is ESI-1, indicating a critical condition, while the model predicts ESI-3, corresponding to a moderate acuity. This type of error is examined in detail because it is closely related to the pronounced class imbalance in the dataset, where mid-acuity visits are substantially more common than critical cases (ESI-3: 5,850 vs. ESI-1: 1,170 in the training set), a pattern that mirrors real-world emergency depart-

ment visit distributions. As a result, models may be biased toward more frequent acuity levels, leading to underestimation of rare but clinically critical cases.

E Additional Experimental Details

E.1 Baselines

We evaluate SOAPTriage against a diverse set of state-of-the-art baselines, including prompting-based LLM approaches, multi-agent frameworks, and encoder-based classification models.

- **Standard Prompting** (Xu et al., 2025) directly queries large language models with task-specific instructions, serving as a strong and widely adopted baseline for clinical reasoning tasks.
- **Prompt-RAG** (Lewis et al., 2020) augments standard prompting with retrieval-augmented generation by retrieving relevant clinical triage knowledge from an external corpus and injecting it into the prompt as additional context, thereby improving factual grounding and robustness for triage prediction.
- **Chain-of-Thought (CoT)** (Kojima et al., 2022) augments prompts with explicit step-by-step reasoning, encouraging LLMs to generate intermediate rationales prior to producing final predictions.
- **Self-Consistency (SCons)** (Wang et al., 2023) extends Chain-of-Thought prompting by sampling multiple reasoning chains and selecting the most consistent outcome among them.
- **Self-Contrast (SCtr)** (Wang et al., 2024) improves robustness by generating multiple reasoning perspectives and reconciling their differences to derive a final decision.
- **Exchange-of-Thought (EoT)** (Yin et al., 2023) facilitates cross-model interaction by exchanging intermediate reasoning traces, enabling complementary reasoning processes to be integrated.
- **Knowledge-Evolvable Assistant (KEA)** (Lu et al., 2025) is a general framework for augmenting large language models. We adopt this framework to construct two specialized repositories based on the MIMIC-IV-ED dataset to support the triage prediction process: (i) an experience bank that stores validated, successful CCPG cases for analogy-based reasoning, and (ii) a reflection bank that records previously misclassified cases along with their corrections and self-summarized error analyses.
- **Task Adaptation and Instruction Tuning (TAIT-LoRA)** (Shen et al., 2025) utilizes instruction tuning for task adaptation. We adopt its instruction-tuning component and perform LoRA-based fine-tuning using the natural language data constructed in our work.
- **TRIAGEAGENT** (Lu et al., 2024) is a multi-agent triage framework that aggregates multiple reasoning perspectives with dynamically updated confidence scores and external evidence, serving as a strong baseline for collaborative clinical decision-making.
- **BERT** (Devlin et al., 2019) is a general-purpose transformer encoder that we fine-tune for ESI classification from triage notes.
- **TCM-BERT** (Yao et al., 2019) and **BioBERT** (Lee et al., 2020) are domain-specific variants of BERT pre-trained on large-scale medical and biomedical corpora, included to assess the effectiveness of specialized medical representations for ESI prediction.
- **KATE-BERT** (Ivanov et al., 2021) extracts medical entities from clinical text and leverages the resulting structured representations for downstream triage classification.

E.2 Implementation and Experimental Setup

Implementation Details. We implement SOAPTriage using *Qwen3-8B* as the backbone large language model (LLM). The backbone consists of 36 transformer layers. In the SGE module, we extract hidden representations from intermediate layers 18-22, taking one representation from each selected layer. For each layer, mean pooling is applied across the sequence dimension, followed by L2 normalization, yielding a set of stream-level embeddings.

To fuse information from multiple streams, we employ a lightweight two-layer multilayer perceptron (MLP) as the gating network. The gating network operates on the concatenation of all stream embeddings and produces a normalized weight for

each stream. These weights are then used to compute a weighted sum of the stream representations, resulting in a single fused embedding, which is subsequently fed into the downstream prediction head for inference.

For ordinal prediction, we apply a sigmoid function to the threshold logits and determine the ordinal class by counting the number of thresholds exceeding 0.5. To mitigate class imbalance, we apply an adaptive class-weighting strategy during training, automatically assigning larger weights to rarer classes and smaller weights to more frequent ones, thereby increasing the model's sensitivity to minority-class errors.

All MIMIC-derived datasets used in this work are accessed and processed in accordance with the official MIMIC-IV data use agreement and license. The processed data released with this work contain no protected health information and comply with the original dataset terms. We do not collect, infer, or redistribute any personally identifying information, and all generated or processed data preserve the anonymization guarantees of the original sources.

Training Configuration. All models are trained using the AdamW optimizer with a weight decay of 5×10^{-4} and an initial learning rate of 3×10^{-4}. We adopt a cosine annealing schedule to adjust the learning rate throughout training. Training is conducted with a batch size of 128 for 200 epochs, and gradient clipping with a maximum norm of 5.0 is applied to stabilize optimization.

For frozen LLM components, we set the decoding temperature to 0 to ensure deterministic behavior. Model selection is based on validation performance: we select the checkpoint achieving the best validation results and additionally retain the top three checkpoints to enhance robustness.

Retrieval and Experimental Setup. During retrieval in the CNA module, we use *bge-large-en-v1.5* as the encoder model and set the retrieval top-K to 3. All experiments are conducted on a single NVIDIA RTX 5090 GPU. To ensure reproducibility, we fix the random seed to 42 across all experiments.

F Additional Experiments

To further validate the robustness and generalizability of SOAPTriage beyond the MIMIC-IV benchmark, we conduct additional experiments on the NHAMCS dataset and a real-world clinical-notes dataset (Lu et al., 2024). For NHAMCS, we follow the same training–validation–test protocol as in MIMIC-IV and re-train all methods on NHAMCS to assess in-domain performance under a shifted data distribution. In contrast, to approximate a realistic deployment scenario, we evaluate cross-dataset transfer by directly applying the model trained on MIMIC-IV to 384 real-world clinical notes without any further fine-tuning.

F.1 Experiments on NHAMCS Dataset

We evaluate SOAPTriage and representative baseline methods on the NHAMCS dataset, with results for automated ESI prediction summarized in Table 9. Consistent with our primary experiments on MIMIC-IV, encoder-based methods (e.g., BERT and TCM-BERT) generally outperform prompting-based approaches such as Chain-of-Thought (CoT) prompting, as well as multi-agent frameworks like TRIAGEAGENT. Across most evaluation metrics, SOAPTriage consistently outperforms all compared baselines. While we observe a marginal increase in UnderTriage relative to certain methods, likely attributable to the imbalanced class distribution in NHAMCS, our approach achieves substantial improvements in other critical indicators. In particular, it significantly reduces Total Discordance, indicating a more balanced and reliable overall triage performance.

F.2 Experiments on Real-world Clinical Notes

To assess performance in a realistic clinical setting, we further evaluate our approach on an independent dataset consisting of 384 real-world clinical notes collected from an actual clinical environment. In this setting, the model trained on MIMIC-IV is directly applied to the real-world dataset without any additional fine-tuning, enabling an evaluation of cross-dataset generalization. We compare SOAPTriage against representative baselines, including Chain-of-Thought (CoT), Self-Contrast (SCtr), Self-Consistency (SCons), and BERT. As shown in Table 8, although SOAPTriage does not achieve the lowest Total Discordance, SOAPTriage outperforms CoT, SCtr, and BERT in Total Discordance, while avoiding the pronounced over-triage tendency observed in BERT. These findings indicate that the performance gains of SOAPTriage are driven by improved ESI reasoning ability, rather than simply by adaptation to the synthetic note style introduced by CNA.

Model	Total ↓	Under ↓	S-Under ↓	Over ↓	S-Over ↓
CoT (1-Agent)	64.06	28.90	14.06	35.15	0.52
SCtr (1-Agent)	56.51	24.73	16.14	31.77	**0.26**
SCons (5-Agent)	**48.95**	29.68	**12.23**	**19.27**	0.52
BERT	71.35	**7.55**	3.64	63.80	17.18
SOAPTriage (8B)	52.60	20.57	10.93	32.03	1.56

Table 8: Performance comparison (%) of different methods across five triage metrics. We report Total Discordance, UnderTriage, and OverTriage rates, as well as their clinically significant counterparts (*S-Under* and *S-Over*). Lower values indicate better performance.

G Prompt

Figure 10 illustrates the prompt templates used in SOAPTriage to guide the language model to extract Subjective, Objective, Assessment, and Plan information from each ED note.

Method	Total Discordance↓	UnderTriage↓	Significant UnderTriage↓	OverTriage↓	Significant OverTriage↓
Prompt-based					
Prompt	63.86	17.76	<u>7.85</u>	46.10	1.75
Prompt-RAG	58.05	16.58	**6.40**	41.46	1.15
CoT (1-Agent)	54.86	16.43	10.15	38.42	1.75
SCons (5-Agent)	55.77	**13.11**	9.12	42.65	1.33
SCtr (1-Agent)	57.65	22.09	15.77	35.54	1.90
EoT (3-Agent)	54.50	<u>14.68</u>	10.21	39.81	1.45
Multi-agent Methods					
TRIAGEAGENT	52.63	21.57	9.90	31.05	1.32
Encoder-based					
BERT	46.28	20.84	10.75	25.43	<u>0.33</u>
TCM-BERT	46.46	23.87	13.47	22.60	0.42
BioBERT	<u>46.22</u>	24.71	11.24	<u>21.51</u>	<u>0.33</u>
SOAPTriage (8B)	**44.77**	25.86	15.46	**18.91**	**0.30**

Table 9: Performance comparison (%) between SOAPTriage and baseline methods on the NHAMCS dataset. Best results are highlighted in bold, while the second-best results are indicated by underlining. Lower values indicate better performance.

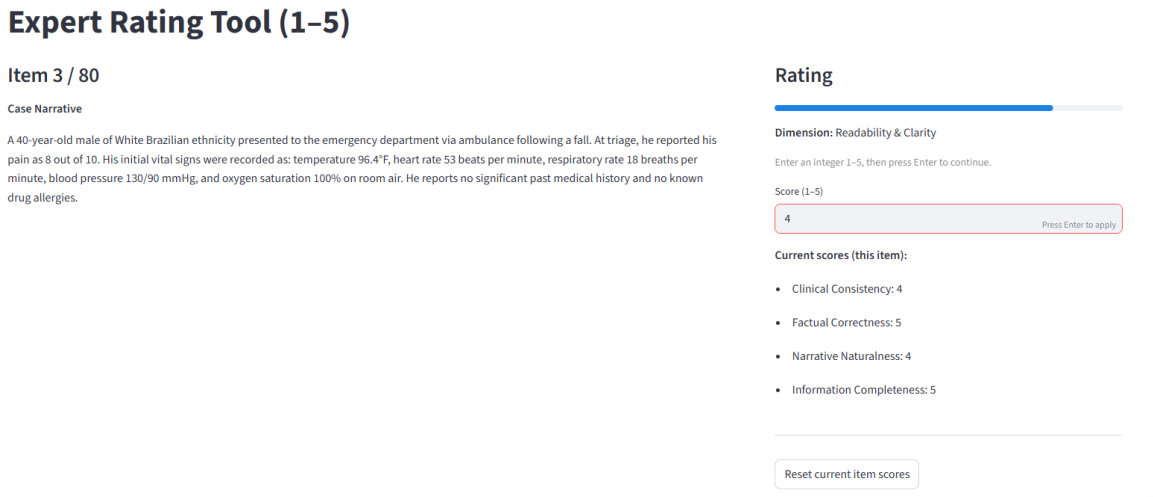

Figure 7: Annotation interface illustrating five evaluation dimensions: clinical consistency, factual correctness, narrative naturalness, information completeness, and readability and clarity.

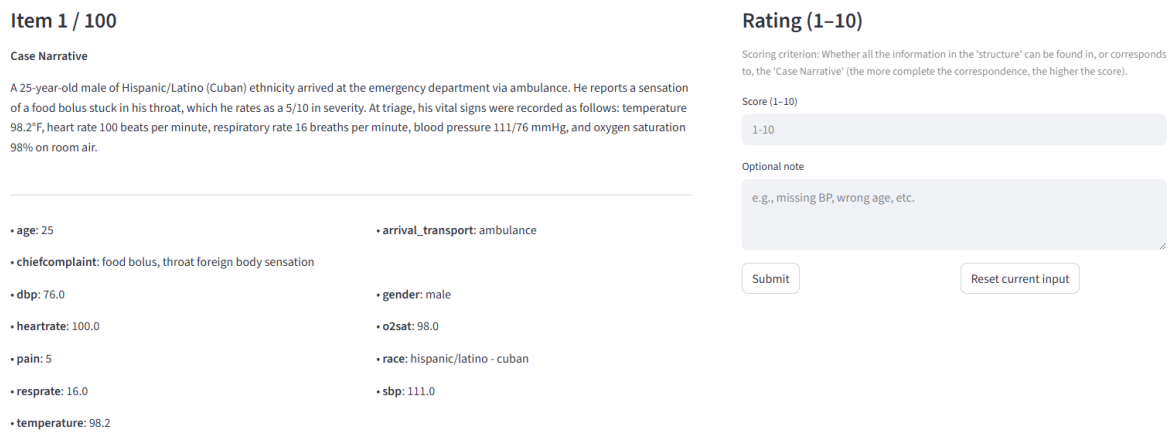

Figure 8: Annotation interface for assessing hallucination-related errors.

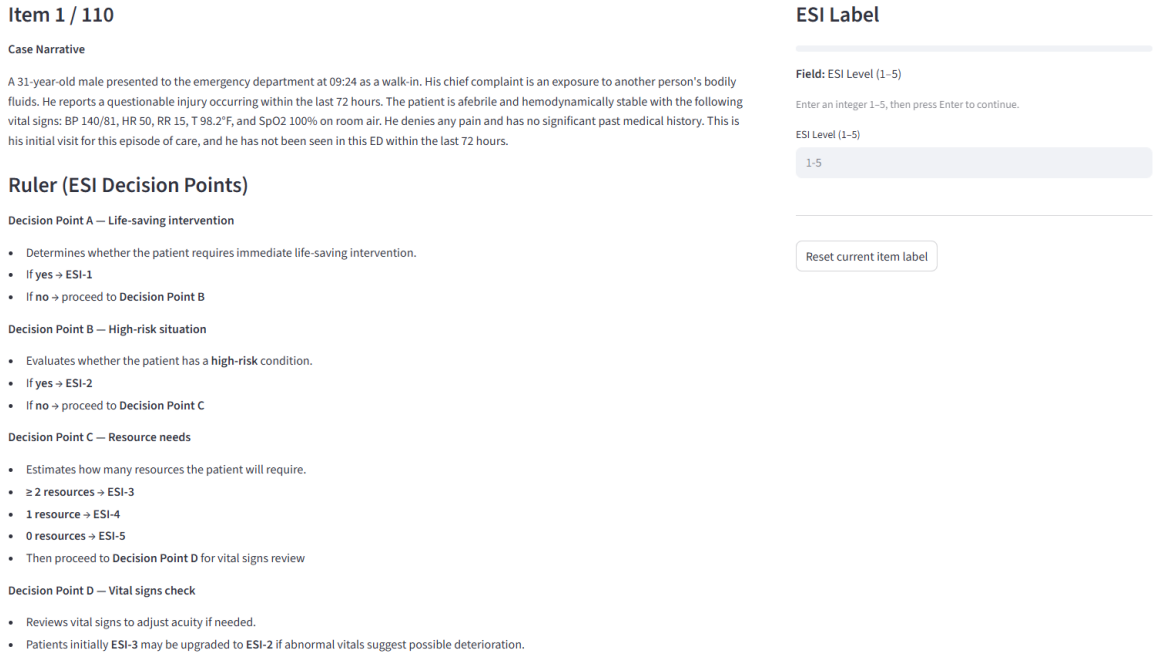

Figure 9: Annotation interface for annotating ESI triage levels.

Prompt Template used for SOAP-guided Encoding

System Prompt:
You are an emergency triage doctor using the SOAP framework for triage (levels 1-5, with 1 being most urgent).

Input:
Below is a short triage note describing a patient's condition: {text}

Subjective Task:
1. Extract and analyze only the SUBJECTIVE: what the patient or caregiver reports in their own words, including main complaint, onset, context, and perceived severity.
2. Ignore physical exam findings, vital signs, measurements, and any management plan.
3. Write 2-3 concise sentences in English, focusing purely on the subjective complaint.
Output ONLY the analyzed subjective description, without any labels or extra commentary.

Objective Task:
1. Extract and analyze only the OBJECTIVE: vital signs, observable findings, measurements, and documented exam facts.
2. Do NOT include patient-reported symptoms unless they are clearly measured/observed.
3. Write 2-3 concise sentences in English.
Output ONLY the rewritten objective description, with no labels or extra commentary.

Assessment Task:
1. Based on the text, infer and summarize the clinical ASSESSMENT: likely problem(s), possible differential diagnoses, and overall severity or risk.
2. Focus on short phrases like "minor soft tissue injury", "possible head trauma", "suspected asthma exacerbation", "high risk chest pain", etc.
3. Use 2-3 concise sentences in English, emphasizing severity and risk (e.g., stable vs. potentially life-threatening), but DO NOT mention any specific triage level number.
Output ONLY the rewritten assessment, without any labels or extra commentary.

Plan Task:
1. Infer and summarize the management PLAN (P in SOAP) that would reasonably follow from this triage note in an emergency department.
2. Focus on the types of resources and actions likely needed: examples include wound cleansing and suturing, observation, laboratory tests, imaging studies, medications, monitoring of vital signs, or immediate life-saving interventions.
3. Use 1-3 concise sentences in English describing the recommended next steps and resources, but DO NOT mention any specific triage level number.
Output ONLY the inferred management plan, without any labels or extra commentary.

Figure 10: Prompt template used for SOAP-guided encoding in the SOAPTriage framework.