

Towards glass-box triage: an interpretable multi-agent clinical decision support system for emergency department

Roxana Ramona Szomiu, Darius Robert Ember, Adrian Groza

Artificial Intelligence Research Institute AIRi@UTCN

Department of Computer Science

Technical University of Cluj-Napoca

Cluj-Napoca, Romania

Email: Roxana.Szomiu@cs.utcluj.ro, Ember.Gh.Darius@student.utcluj.ro, Adrian.Groza@cs.utcluj.ro

Abstract—Emergency Departments (EDs) are high-pressure, resource-constrained environments in which accurate and rapid patient triage is critical to clinical safety. The patient’s initial assessment determines the urgency of care required and shapes their subsequent clinical pathway - including resource allocation, specialist routing, and final disposition. While machine learning (ML) have demonstrated strong predictive capabilities for clinical outcomes, their adoption in real-world healthcare settings remains limited by a lack of interpretability: such models typically operate as black boxes, returning probabilities without transparent clinical reasoning.

To address this gap, we introduce a multi-agent, pipeline-based clinical decision support system for ED triage. Rather than relying on a single model, the architecture delegates distinct clinical tasks to specialized agents, using Large Language Models (LLMs) exclusively to parse patient narratives and manage the interaction with patient, while all clinical judgments (acuity scoring, diagnosis, and hospital disposition) are handled by calibrated, supervised XGBoost classifiers. Trained and validated on over 400,000 real-world encounters from the MIMIC-IV-ED dataset, the system also distinguishes between an initial assessment and a refined assessment based on supplementary nurse-collected data, quantifying the contribution of additional clinical information to the triage decision.

This paper presents the system’s architecture and evaluates its clinical safety with a focus on under-triage mitigation, showing that this decoupled design preserves full decision traceability without sacrificing predictive performance.

Index Terms—clinical triage, machine learning, large language models, multi-agent system, specialized agents, emergency department, evidence-based nursing, clinical decision support, interpretability

I. INTRODUCTION

Emergency departments (EDs) across Europe and beyond are experiencing growing pressure from increasing demand and limited resources. Across 26 OECD¹ countries, EDs recorded an average of 31 visits per 100 population in 2023, with approximately 20% of these visits resulting in hospital admission. This demand differs across countries: annual ED utilization ranges from fewer than 15 visits per 100 population in some health systems to more than 65 visits per 100

population in countries such as Portugal and Spain [1]. At the population level, the European Society for Emergency Medicine (EUSEM) estimates that approximately 25% of Europeans visit an ED each year - equivalent to 100–300 visits per 1,000 inhabitants annually depending on the country - and that 70–80% of all hospital admissions originate from the ED [2]. Together, these statistics demonstrate that the ED is a central gateway for acute care triage rather than a peripheral entry point to the healthcare system.

Triage - the process of prioritizing patients based on the urgency of their condition at arrival - is a key component of emergency care. Accurate triage is essential: under-triage can delay care for critically ill patients, while over-triage can waste limited healthcare resources.

Over recent decades, EDs have experienced increasing patient volumes, especially among more complex and critically ill cases, while emergency care resources have remained limited. This imbalance has contributed to ongoing overcrowding and longer waits before patients receive triage and treatment. Since, triage is the first step in emergency care, delays at this stage can affect the entire patient pathway.

These challenges highlight the need for improved triage strategies that optimize resource use and support clinical decision-making. Based on these limitations, AI-based systems have emerged as a promising approach to enhance ED triage. Recent progress in ML and LLMs has the potential to improve clinical decision support, reduce variability in triage assessments, and enable more efficient patient prioritization under high-demand conditions.

Although traditional predictive models often achieve high accuracy, they typically operate as black boxes, providing little insight into the reasoning behind their predictions. This lack of interpretability limits the use of clinical applications in the real-world, as medical experts must be able to understand and justify decisions that directly affect patient care.

To address this limitation, we propose a modular, pipeline-based framework that mirrors the sequential clinical triage workflow. Using specialized LLM-based agents, the system separates natural language understanding from structured ML-

¹Organisation for Economic Co-operation and Development

based clinical prediction, enabling a more transparent and interpretable decision-making process.

Our system exposes the reasoning process at each stage of the clinical workflow, allowing clinicians to inspect, verify, and interpret intermediate decisions. By combining modular reasoning with transparent decision support, the system improves accountability, increases clinician trust, and supports—rather than replaces—human judgment in emergency triage.

II. RELATED WORK

In recent years, the use of LLMs has increased significantly across various domains. In parallel, the healthcare systems have evolved from single-task predictive models to complex, collaborative multi-agent systems (MAS). Recent literature further highlights the potential of LLMs to support and simulate aspects of clinical reasoning through automated and interactive decision-making frameworks. Research on automated support for medical triage has evolved along two largely parallel tracks: (1) classical machine learning (ML) and deep learning (DL) models trained on structured/unstructured EHR data, and (2) more recent evaluations of LLMs applied zero-shot or few-shot to triage decision-making.

Supervised ML and DL models trained on initial triage data have consistently outperformed traditional rule-based systems such as the Emergency Severity Index (ESI) protocol [3]. Raita et al. [4] showed that gradient-boosted and deep neural network models predicted hospitalization and critical-care outcomes more accurately than an ESI-based logistic regression baseline while reducing over-triage. Yao et al. [5] further demonstrated that combining structured vital signs with unstructured electronic health record (EHR) data improves outcome prediction beyond ESI alone. Extending this work to clinical practice, Vântu et al. [6] showed that automated ML pipelines applied to raw triage data can support scalable and consistent decision-making in hospital settings. A systematic review by Porto [7] concluded that ML and NLP approaches generally improve triage accuracy and consistency compared with conventional methods, while emphasizing the need for stronger external validation and broader clinical implementation.

While classical ML/DL models rely on task-specific training, recent research has explored zero-shot LLMs (e.g., GPT-4 model) as flexible triage decision-support systems that require no retraining. Studies [8], [9], [10] found that LLMs can achieve competitive, sometimes physician-comparable, triage performance through language-based reasoning, but they become less reliable for complex cases and remain unsuitable for autonomous decision-making. These findings are further supported by Cui et al. [11] in a systematic review and meta-analysis that reported high specificity (97%) but only moderate sensitivity (61%) to identify patients with the highest-acuity, suggesting strong ability to rule out lower-risk cases, but limitations in detecting the most urgent patients, while Lee et al. [12] found that LLM performance differed across demographic groups. Overall, current evidence supports zero-shot LLMs as

clinician decision-support tools rather than autonomous triage systems.

Recent studies have explored the use of multi-agent LLM architectures for clinical tasks beyond single-model prediction, revealing that the benefits of agentic collaboration are task-dependent, with improvements over baseline LLMs observed in specific settings.

TriageAgent [13] specifically targets clinical triage through a dynamic multi-agent framework in which LLM-based agents are engaged in multi-round discussion, verify each other’s reasoning, incorporate self-confidence and early-stopping mechanisms, and ground their decisions using retrieval-augmented generation over the ESI handbook. The framework achieves higher diagnostic and triage accuracy than state-of-the-art single-LLM baselines across three clinical triage benchmarks, suggesting that structured multi-agent collaboration reduces isolated hallucinations and improves triage performance.

The authors [14] introduced the MAP framework that combines specialized triage, diagnosis, and treatment agents under a supervising agent, achieving a 25.1% improvement in diagnostic accuracy over HuatuoGPT2-13B and outperforming clinicians on the IPDS benchmark of 51,274 MIMIC-IV cases. Together, *TriageAgent* and MAP demonstrate the potential of multi-agent LLM collaboration to surpass single-model baselines in clinical prediction.

Chen et al. [15] conducted the first large-scale systematic review and meta-analysis comparing the independent and collaborative performance of LLMs and clinicians in diagnosis and triage. Across 50 studies evaluating 25 LLMs, they found that LLMs achieved triage accuracy comparable to clinicians, while clinician-LLM collaboration consistently improved diagnostic performance over clinicians alone. The results support the integration of LLMs as decision-support tools in clinical triage rather than autonomous replacements for clinicians.

Several publicly available datasets are widely used in the literature for ED triage and downstream clinical prediction addressing complementary aspects, including MIMIC-IV-ED [16] and eICU-CRD [17]. MIMIC-IV-ED supports emergency department triage by capturing patient encounters from arrival through disposition, whereas eICU-CRD focuses on post-triage critical care and is primarily used for outcome prediction and cross-institutional validation.

III. METHODOLOGY

The proposed methodology follows the multi-agent system paradigm by breaking the main task into specialized sub-tasks, each assigned to an agent with a specific responsibility. Instead of relying on a single end-to-end neural network to map a raw patient case directly to a diagnosis, the architecture distributes the processing across multiple agents.

Our system sequentially predicts five clinical outcomes: (i) the Emergency Severity Index (ESI) triage level, (ii) the admission or discharge decision, (iii) the diagnosis for admitted patients, (iv) the appropriate hospital department and (v) the corresponding diagnostic ICD (International Classification of Diseases) code. This separation of responsibilities enables a

predictable transformation of data-from unstructured natural language to structured clinical features, and finally to deterministic clinical predictions - while enhancing modularity and explainability.

A. Dataset: MIMIC-IV-ED

To ground the proposed system in real-world clinical practice, we trained and evaluated it using the MIMIC-IV Emergency Department (MIMIC-IV-ED) dataset [16], a large, publicly available, database containing approximately 425,000 ED encounters collected between 2011 - 2019. We selected this dataset because it is well suited for modeling sequential clinical workflow in ED triage. It includes triage assessments, longitudinal vital signs, medication reconciliation records, discharge diagnoses, and ED stay information, providing a comprehensive representation of the patient care pathway.

The dataset is organized into multiple tables capturing different stages of the ED visit: triage (chief complaint, pain score, and initial vital signs), vitalsign (timestamped vital sign measurements), medrecon (pre-admission medications), diagnosis (final ICD-9/ICD-10 diagnoses), and edstays (arrival and discharge times, length of stay, and patient disposition).

The timestamped structure enables realistic modeling of sequential clinical workflows by restricting each predictive model to information available at the corresponding stage of care. For example, the first agent of the pipeline (Triage Agent) is trained solely on initial triage data, whereas next agents incorporate longitudinal vital signs and medication histories collected later during the ED stay. This temporal separation prevents data leakage and ensures clinically valid model development.

B. System architecture

The architecture of our multi-agent system is illustrated in Figure 1. Our solution consists of four specialized agents executing six sequential tasks, coordinated through a deterministic sequential workflow, with each agent responsible for a specific stage of the clinical process. Information is passed sequentially, following the natural progression of the patient’s care pathway:

- 1) **NLP Parser Agent:** is a LLM-based agent that extracts structured features from free text.
- 2) **Triage Agent (ML):** predicts initial ESI-score for acuity and disposition based strictly on intake data using XGBoost.
- 3) **Assistant (Nurse) Agent:** simulates interactive clinical data collection, gathering vitals and information about medications.
- 4) **Doctor Agent (ML Cascade):** refines predictions and establishes a final diagnosis using all available longitudinal data ².

The proposed system is built on the CrewAI orchestration framework, integrating a set of specialized LLM-based agents with dedicated XGBoost predictive models. The architecture

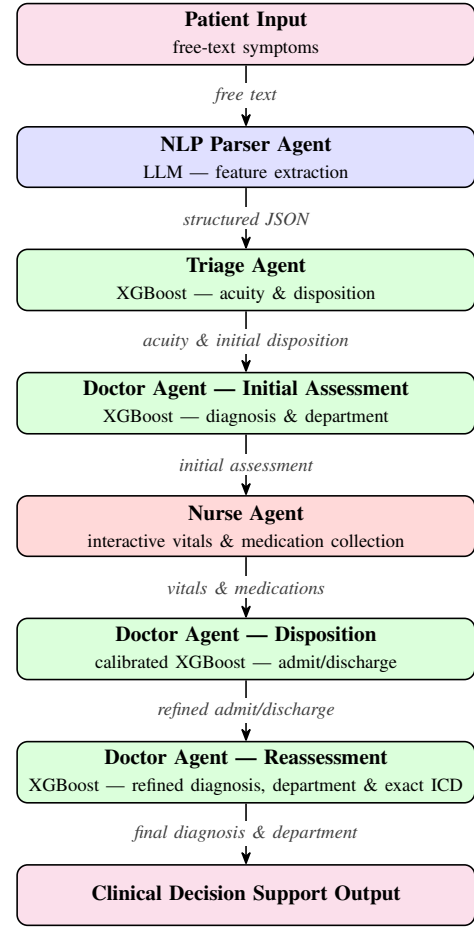


Fig. 1. The sequential pipeline architecture of the clinical prototype. Colour encodes the agent type: LLM-based natural-language processing (blue), supervised machine-learning prediction (green), and interactive data collection (red).

follows a hybrid design principle that explicitly separates two concerns: (i) natural language understanding, handled by the LLM layer, and (ii) clinical prediction, handled by gradient-boosted tree models trained on structured and text-derived features. The LLM does not generate clinical judgments directly; it converts free-text inputs into a structured format, invokes prediction tools that wrap the trained models, and reports their outputs. This design separates language understanding and orchestration from clinical prediction, which remains the responsibility of validated and auditable statistical models.

Information is exchanged between tasks through a lightweight harness layer rather than direct agent-to-agent communication. The harness enforces a structured interaction contract: an upstream agent must produce outputs in a pre-defined format, and downstream agents receive exactly this validated structure [18]. In our system, inter-agent communication relies on two mechanisms: a *context channel* that forwards textual outputs to preserve the clinical narrative, and a *structured-data contract* that appends machine-readable blocks containing key variables (e.g., ESI level, disposition, age, and gender). This fixed interface enables deterministic

²health information collected from the same person over an extended period.

parsing of prediction outputs while keeping agents decoupled from internal prompts, reasoning traces, and intermediate representations.

Each specialized agent processes the output of the previous stage, transforming free-text input into structured clinical predictions. The four agents are invoked across six stages and the Doctor agent runs three times (initial assessment, disposition, and reassessment).

1) *Natural Language Parsing Agent*: The system processes a patient case through a cascading, multi-stage pipeline. Similar to an ED, the workflow begins when a patient arrives and provides a free-text description of their symptoms to the intake module. This LLM-based agent (Gemini 2.5 Flash, MedGemma) is the entry point of the system. The agent transforms unstructured patient-reported text into a structured intake object with a fixed schema containing key demographic, clinical, and arrival information, including: *chief complaints*, *pain score*, *history*, *prior admissions*, *transport mode*, and *available pre-hospital vital signs*.

Missing required fields are resolved through interactive clarification using *AskPatientTool*, rather than imputation, and the parsed intake is verified by a human before initiating triage prediction. This step is crucial because the downstream XGBoost machine learning models require a consistent feature matrix to perform accurately. By isolating the LLM at this specific stage, we completely eliminate the risk of the language model "hallucinating" a diagnosis; its sole, verifiable purpose is data normalization and extraction.

2) *Triage Agent*: Once the patient data is structured into a clean JSON format, it is passed to the Triage Agent. This agent predicts the ESI acuity and admission probability via *TriagePredictionTool*, which wraps two pre-trained XGBoost models. Its input schema uses explicit missingness conventions (e.g., unknown vitals encoded as -1) so the model can represent absence of information rather than have the LLM generated values.

Since the input consists mainly of short chief complaints and demographic features, textual information is transformed using TF-IDF to generate a sparse feature representation that captures the importance of clinically relevant terms. XGBoost was selected for its strong performance on tabular data, efficient handling of sparse inputs, and fast inference. To address clinical class imbalance and reduce harmful under-triage errors, the training objective incorporates class weighting and additional penalties for misclassifying high-severity patients into lower-risk categories, prioritizing patient safety over unweighted accuracy.

3) *Nurse Agent*: Following the initial triage, a patient in a real ED would interact with a triage nurse to have their vital signs measured and their medical history recorded. The system simulates this critical data-gathering phase via the Nurse Agent.

Unlike the other agents, this agent does not rely on a predictive model. Instead, it acquires clinical information through an interactive dialogue, collecting vital signs, timestamped

measurements, medication use, and medical history, while preserving missing values.

A preprocessing pipeline transforms these inputs into the final feature representation, including vital-sign trends, abnormality indicators, medication categories, and history features, using the same processing steps during training and inference. Raw numerical inputs are constrained to biologically plausible ranges (e.g., heart rate: 20–250 bpm) to prevent erroneous values from corrupting the feature vector and affecting downstream predictions.

To ensure robust downstream predictions, the module applies automated data preprocessing. Physiological values are constrained within biologically plausible ranges to prevent erroneous inputs from affecting the model.

When the LLM parses patient responses, any information that is unknown or not provided by the patient is simply marked as missing (e.g., using -1, 'None', or 'NaN') rather than being estimated. To remain consistent with the way the models were trained, the Doctor Disposition, Nurse, and Triage tools then handle these missing values using median imputation based on the training set, together with missingness indicators that allow the models to capture the information carried by the absence of a measurement. In addition, free-text medication entries are grouped into broad clinical categories, such as cardiovascular, anticoagulant, and respiratory medications, providing useful information about the patient's medical history without requiring direct access to electronic health records (EHRs).

At runtime, the nurse agent extracts the required clinical variables by asking the patient targeted questions using an LLM. Information already available in the patient record (e.g., past medical history, current medications, and other structured data) is retrieved through dedicated tools that provide access to the corresponding dataset fields. The agent is prompted to collect all features required by the downstream prediction model; when a required variable is unavailable from the dataset, it is requested directly from the patient. However, patients are not required to answer every question, and missing responses (e.g., declining to provide information about current medications) are handled as unavailable inputs.

4) *Doctor Agent*: The Doctor Agent represents the final decision-making component of the pipeline. It integrates the complete patient representation - including the NLP-parsed intake information, the predicted ESI level from the Triage agent, and the validated clinical data collected by the Nurse agent - to perform a multi-stage diagnostic inference process.

The Doctor agent is based on three different XGBoost trained models for each type of task. The first is a *diagnosis model*, which predicts the patient's primary diagnostic category. The second is a *department model*, responsible for recommending the appropriate hospital department. The third is a *disposition model*, which estimates the probability of hospital admission based on the complete clinical picture. To ensure that the predicted probabilities accurately reflect observed admission rates, the disposition model is calibrated

using isotonic regression, producing well-calibrated risk estimates suitable for clinical decision support.

The diagnostic workflow is explicitly divided into: stage 1 (initial assessment) and stage 2 (reassessment). During the stage 1, only the information available at triage is considered to generate an initial prediction of the patient’s primary diagnostic category (e.g., cardiovascular, respiratory, or gastrointestinal) together with the most appropriate hospital department for patient routing. This stage provides an early diagnostic hypothesis before incorporating additional clinical information obtained during the subsequent reassessment. Stage 2 functions as an advisory module that refines this prediction by estimating the most likely ICD code within the Stage 1 category, without altering the predicted diagnostic category itself.

IV. EXPERIMENTS AND RESULTS

A. Evaluation protocol

The evaluation is conducted at two levels. First, a quantitative evaluation assesses each model using feature vectors derived directly from the structured fields of the dataset across multiple test cases. Secondly, while the tabular evaluation uses structured feature vectors, real-world deployment requires an LLM to extract these features from free-text patient descriptions. To assess the impact of this parsing step, we performed an end-to-end evaluation on 250 synthetic natural-language cases derived from real emergency department presentations.

By comparing four input configurations that share the same prediction models but differ only in how the patient’s complaint reaches them, we isolate the error introduced by the LLM parser from the error of the predictive models themselves. This breakdown is reported in Section IV-D and Table IV.

Specifically, the tabular sets used for testing contain 83,617 encounters to predict acuity and disposition and 20,420 admitted encounters to predict the diagnosis, department and ICD code (80/20 stratified split). The 250 generated cases are split into 163 admitted and 87 discharged patients. The *filtered feature vector* and *direct invocation* both use the exact tabular data and never invoke the LLM, whereas the *Gemini Flash* and *MedGemma* based parser use the information extracted by the LLM from unstructured input. Consequently, the isolated impact of the parsing stage can be quantified as the performance difference between direct invocation and the corresponding parser-based configuration.

B. Tabular results on the held-out test sets

Table I summarizes the performance on the large test sets. For acuity prediction (ESI 1–5), on the 83,617 test cases, the model achieved 67.55% exact accuracy, with 98.18% of predictions within one ESI level of the ground truth. Given the ordinal nature of the ESI scale, the model also achieved a quadratically weighted Cohen’s K of 0.64 (using *cohen_kappa_score* from scikit).

For diagnosis and department, the table shows the Doctor’s initial assessment (triage data only) and its reassessment

after the vital signs, medication and past medical history are collected by the Nurse agent, emphasizing the value of the additional data collected.

Target (classes)	<i>n</i>	Headline metrics
ESI acuity (5)	83,617	exact 67.6%, within-1 98.2%, QWK 0.64
Disposition (2, calibrated)	83,617	acc 83.3%, AUC 0.914
Diagnosis (13), initial	20,420	top-1 60.1%, top-3 82.7%
Diagnosis (13), post-nurse	20,420	top-1 64.1% , top-3 85.7%
Department (11), initial	20,420	top-1 60.7%, top-3 92.3%
Department (11), post-nurse	20,420	top-1 70.8% , top-3 94.1%
Exact ICD (3-char, Stage-2)	20,420	union 64.9%, graded 0.90

TABLE I
TABULAR PERFORMANCE OF EACH PREDICTION STAGE ON TEST SETS.
QWK: QUADRATIC-WEIGHTED COHEN’S κ

The collected data brings a significant improvement in diagnosis in Top-1 from 60.1% to 64.1% and a large Top-1 gain in department from 60.7% to 70.8%. We have to mention that the department gain comes from a higher concentration around the majority class, meaning *General Medicine* department recall rises from 0.65 to 0.87 while some minority departments decrease, so the balanced accuracy drops 4.4pp, from 49.6% to 45.2%, as Top-1 rises, from 60.7% to 70.8%.

For diagnosis, the improvement is consistent across all classes. For exact ICD code prediction, however, strict top-5 recall does not fully capture the model’s clinical value. This metric is calculated only within the correct (oracle) ICD category. The Stage-2 model ranks all possible exact codes in that category, and a prediction is counted as correct only if the true 3-character ICD code appears in the top five suggestions. This makes it a very strict, all-or-nothing measure. A prediction that is clinically very close to the correct code, such as a neighboring code in the same ICD branch, receives the same score of zero as a completely unrelated prediction.

To account for these borderline cases, we also use three graded evaluation measures: cosine similarity between the ICD titles, cosine similarity between the title embeddings, and distance within the ICD hierarchy. For each case, these measures assign the highest similarity score between the true code and any of the top five predicted codes. Because an exact match always receives the maximum score, the graded metrics are always equal to or higher than the strict recall score. The difference between them represents the credit given to predictions that are clinically similar, even if they are not exact matches.

For example, within the correct ICD category, the strict top-5 recall is 0.69. However, the graded top-5 score based on title-embedding similarity increases to 0.91. This indicates that most of the apparent errors are actually predictions of codes that are very close to the correct one.

We performed hyperparameter optimization using the Optuna framework. Optuna’s multi-objective NSGA-II algorithm

was used to optimize the triage models, while separate single-objective optimization procedures were carried out for the doctor models. The results showed that the current manually selected configuration is already close to optimal, and no other configuration was found to perform better overall.

C. Probability calibration and operating points

A clinical decision-support system should not only prioritize patients according to their admission risk but also produce reliable admission probabilities, so that each operating threshold corresponds to the real admission rate. Table II compares the raw triage screening probability with the isotonic-calibrated Doctor disposition model, on the same 83,617-row test set. Calibration greatly lowers the expected calibration error (ECE), roughly twenty times (from 0.0746 to 0.0036), making the reliability curve essentially diagonal, while also improving the model’s ability to distinguish between patients who will and will not be admitted.

The table reports the four evaluation metrics used: AUC (Area Under the Curve), which measures ranking quality (higher values indicate better discrimination); Brier score, which quantifies the mean squared error of predicted probabilities; and ECE/MCE (Expected and Mean Calibration Error), which measure the average and worst-case differences between predicted probabilities and observed admission rates, respectively (lower values indicate better calibration).

TABLE II
DISPOSITION CALIBRATION QUALITY: RAW TRIAGE SCREEN VS. ISOTONIC-CALIBRATED DOCTOR MODEL.

Disposition model	AUC	Brier	ECE	MCE
Triage screen (uncalibrated)	0.864	0.150	0.0746	0.137
Doctor (isotonic-calibrated)	0.914	0.113	0.0036	0.015

Because the calibrated probabilities are trustworthy, the operating threshold becomes an intentional configurable value for clinical safety, rather than a tuning artifact. Table III shows how the lower threshold values lead to worse overall accuracy, but also a much lower under-triage rate (missed admissions). The deployed threshold of 0.40 sacrifices only 0.6pp of accuracy compared to 0.50 while lowering the under-triage rate from 23.5% to 18.1% (sensitivity 0.765 $\rightarrow$ 0.819). The threshold can be lowered even further to reach even stricter safety targets (e.g. 13.5% under-triage at 0.30). This directly addresses the mortality risk associated with under-triage, highlighted in the introduction. The department model is also isotonically calibrated (ECE 0.019), which accounts for most of its post-nurse top-1 gain, though concentrated on the majority class, as we noted before.

The table reports Doctor disposition operating points at the deployed probability threshold of 0.40. Under-triage and over-triage represent missed admissions (1–Sens) and unnecessary admissions (1–Spec), respectively. BA (Balanced Accuracy) is defined as the mean of sensitivity and specificity.

TABLE III
DOCTOR DISPOSITION OPERATING POINTS (CALIBRATED MODEL)

Threshold	Accuracy	Sensitivity	Specificity	Under	Over	BA
0.15	0.750	0.946	0.636	0.054	0.364	0.791
0.20	0.782	0.924	0.699	0.076	0.301	0.811
0.30	0.820	0.865	0.793	0.135	0.207	0.829
0.40	0.833	0.819	0.842	0.181	0.158	0.830
0.50	0.840	0.765	0.883	0.235	0.117	0.824

D. End-to-end evaluation and the cost of the NLP parser

Table IV reports the four configurations on the 250 generated cases. The two LLM-free configurations (filtered feature vector and direct invocation) are almost identical, so the prediction pipeline itself incurs negligible loss. Almost the entire end-to-end degradation comes from the LLM parser. Measured as the drop from direct invocation to the Flash parser, the isolated parser cost is -2.0 pp for acuity and -2.4 pp for disposition, but -6.8 pp for diagnosis, -8.6 pp for the ICD union, and -11.7 pp for department. Acuity and disposition remain relatively robust to parsing errors, whereas tasks that depend heavily on interpreting the patient’s complaints are more sensitive. This is likely because the parser must translate lay descriptions into the abbreviated clinical terminology used in medical records (e.g. converting “brain bleed” into “subdural haemorrhage”), which is a lossy process. We note that the lower end-to-end performance numbers reflect this extraction cost together with the much smaller sample size. Therefore, they are not a direct comparison against the full tabular test sets.

The two LLM-free configurations show comparable performance, indicating that the remaining degradation observed in the parser-based configurations can be attributed to the NLP parser.

TABLE IV
END-TO-END BREAKDOWN ON THE 250 SYNTHETIC CASES (DIAGNOSIS, DEPARTMENT, AND ICD ON THE 163 ADMITTED PATIENTS).

Metric	Feat. vec.	Direct	Flash	MedG.
Acuity, exact accuracy	68.4%	68.0%	66.0%	60.4%
Disposition, acc. @0.40	82.0%	81.6%	79.2%	78.4%
Top-1 diagnosis	47.2%	49.1%	42.3%	42.3%
Top-1 department	58.9%	59.5%	47.9%	49.7%
ICD, candidate union	53.4%	55.2%	46.6%	46.0%

E. Language-model backend comparison

Table V presents the results for the two switchable parser back-ends, the hosted Gemini 2.5 Flash and the self-hosted MedGemma-4B, in the same parser configuration (on 250 cases). Flash more accurately translates patients’ lay descriptions into the abbreviated clinical terminology used in medical records, as reflected by its higher complaint-token Jaccard score. This leads to decisively better ESI acuity and disposition, marginally better exact-ICD, and a nearly unbiased acuity error (mean signed error $+0.008$ vs. MedGemma’s -0.152 , indicating that MedGemma-4B consistently assigns patients to higher acuity levels than necessary). MedGemma wins department top-1 by 1.8pp, and diagnosis is essentially a

tie (0.423/0.626 top-1/top-3). Based on these results, Gemini 2.5 Flash is the deployed backend, but the parser-related performance loss is similar for both backends, indicating that the remaining gap comes from the challenges of parsing free text, rather than from the specific model used.

TABLE V
FLASH VS. MEDGEMMA-4B AS THE NLP PARSE. QWK:
QUADRATIC-WEIGHTED κ ON ACUITY.

Metric	Flash	MedGemma
Complaint-token Jaccard	0.598	0.529
Acuity, exact accuracy	0.660	0.604
Acuity, QWK	0.615	0.539
Disposition, ROC-AUC	0.888	0.873
Department, top-1	0.479	0.497
Exact ICD, union	0.466	0.460

F. An end-to-end running scenario

Fig. 2 presents a real clinical case processed through all six stages of the proposed pipeline, demonstrating its transparent design. Instead of producing only a final decision, each stage yields an interpretable intermediate output, allowing users to trace the reasoning process leading to the final diagnosis. The illustrated case involves an 83-year-old man who is brought to the emergency department by ambulance after a fall, presenting with severe left hip pain.

The parser extracts the structured fields, then Triage agent assigns ESI 3 score and an initial disposition of *admission*. The Doctor agent’s initial assessment classifies the case into the *Injury and Poisoning* category. The Nurse agent collects vital signs and medication history. Then, the calibrated disposition model confirms *admission* with high confidence ($P = 0.98$). The Doctor’s reassessment confirms the diagnosis, sends the patient to the *Orthopedics department*, and returns the exact ICD code *S72 (displaced femur fracture)* as its top candidate, via the advisory Stage-2 resolver. A certified clinician can follow and verify the entire chain of decisions, from “fall” and “hip pain” to a specific fracture code, at each step of the process. For this case, every stage of the pipeline produces the correct output. The predicted ESI level (3), disposition (admitted), specialty (Injury and Poisoning/Orthopedics), and ICD-10 code (S72) all match the ground truth.

```

Patient (free text):
"I'm 83 and I fell, now my left hip is just
killing me, a 10/10. The ambulance brought me."

[1] Parser -> {age:83, sex:M, arrival:ambulance,
complaint:"fall, hip pain",
pain:10/10}
[2] Triage -> ESI 3 (true: 3);
initial disposition: ADMIT
[3] Doctor -> Dx#1: Injury and Poisoning
initial (top-3: Injury, Musculoskel., Resp.)
[4] Nurse -> vitals + medication history
[5] Doctor -> refined P(admit)=0.98 >= 0.40
dispo. => ADMIT (confirms triage screen)
[6] Doctor -> Dx: Injury and Poisoning;
reassess Dept: ORTHO; exact ICD #1: S72
"displaced femur fracture" (correct)

```

Fig. 2. End-to-end trace for a real case (stay 32839233).

V. DISCUSSIONS

The proposed multi-agent framework demonstrates that combining LLMs for natural language understanding with traditional ML models for clinical prediction offers a viable and safe alternative to end-to-end “black box” AI systems in emergency triage. By following the same step-by-step process used in real-world triage, the system delivers strong predictive performance while preserving the level of transparency and interpretability that clinicians need for decision support.

A key barrier to clinical AI adoption is the lack of transparency in model reasoning. Our architecture explicitly addresses this issue by separating information extraction from the statistical prediction model, making the system’s reasoning transparent and easy to evaluate. By relying on LLMs only to parse patient’s unstructured description into a structured schema, rather than generating clinical judgements, we eliminate the risk of hallucinated diagnoses. The pipeline provides verifiable intermediate outputs at each stage, allowing clinicians to identify and correct errors before they propagate. This transparency ensures that final predictions (ESI, disposition, department, diagnosis, and ICD code) are based on traceable clinical features, improving verification and supporting trust in critical emergency care settings.

A. Patient safety through probability calibration

For a clinical support system, raw accuracy is less important than the rate of critical under-triage, where critically-ill patients are assigned a lower urgency level than they should. Our tests showed that raw predictive outputs are often calibrated poorly. Our system applies isotonic calibration to the Doctor Agent’s disposition model, and as a result, the expected calibration error improved twenty-times, from 0.075 to 0.004.

This calibration allows hospitals to intentionally select thresholds based on clinical safety targets, allowing them to minimize under-triage rates, at the cost of lower raw accuracy. This flexibility directly addresses the risks of mortality which are associated with delayed emergency care, proving that the system can be adapted to the risk tolerance of different hospitals.

B. The bottleneck of natural language parsing

The impact of the parsing stage was not consistent across tasks. Predictions primarily driven by physiological information, such as ESI acuity level and patient disposition, remained relatively robust despite the introduction of the parser. In contrast, tasks that depend heavily on accurately interpreting the patient’s chief complaint - such as department assignment and precise ICD code prediction - experienced substantially greater performance degradation, with department prediction alone dropping by more than 10 percentage points. These findings suggest that translating free-text patient descriptions into concise clinical terminology remains an inherently lossy process.

The backend comparison further supports this finding. The hosted Gemini 2.5 Flash model achieved overall better results compared to the self-hosted MedGemma-4B model, largely

because it was more effective at translating patients' natural language into the standardized clinical terminology used in medical records. This superior text normalization is directly correlated to better ESI acuity assignment and disposition predictions, without the over-triage tendency observed with the smaller MedGemma model.

C. Limitations

Although the results are promising, this study has several limitations. First, the models were trained and evaluated on a single database (MIMIC-IV-ED), which reflects the demographic composition and operational practices of a single healthcare system, which may limit its use in other healthcare systems with different protocols.

Second, the end-to-end evaluation was conducted on a relatively small sample of only 250 generated cases. Although these cases were generated based on real encounters in the ED, they can not fully capture the real-world complexity, ambiguity, and conversational dynamics of live interactions between patients and hospital staff.

Diagnosis prediction is performed using 13 ICD-based categories instead of individual ICD codes. The subsequent exact ICD code prediction is intended only as a supportive step and was not validated against clinical outcomes. Similarly, department assignment uses 11 categories created by merging the original 19 hospital departments. While this simplification improves statistical reliability, it may not provide the level of detail required for applications such as routing patients to specific subspecialties.

VI. CONCLUSION AND FUTURE WORK

This paper presented an interpretable, pipeline-based clinical decision support prototype designed for the emergency department. By separating natural language understanding from deterministic ML, the proposed architecture combines the flexibility of LLMs for information extraction with the robustness, transparency, and reproducibility of tabular predictive models.

The system was trained and tested using MIMIC-IV-ED dataset which reflects the challenges of real-world clinical data, including missing observations, irregular sampling, and noisy free-text chief complaints containing abbreviations and spelling variations. The evaluation demonstrated that the modular design enables accurate prediction across multiple clinical tasks, including triage acuity, diagnostic categorization, department recommendation, and hospital disposition. Safety-oriented training strategies, such as class weighting for critical ESI levels and calibrated admission prediction, improve the reliability of clinically important decisions. Furthermore, adding progressively richer clinical information - from initial triage data to longitudinal vital signs, medication history, and past medical history - substantially enhances downstream diagnostic performance.

As a future work, the architecture can be extended to support multimodal inputs, allowing vision-language models to extract structured findings from diagnostic data such as ECG or EKG.

These features can then be incorporated into the deterministic prediction models to enhance diagnostic performance.

Our framework relies on modeling the sequential transition from an initial triage assessment to a refined assessment after supplementary data collection, requiring datasets that preserve the arrival-to-decision workflow. Among the available public datasets, only MIMIC-IV-ED provides this structure, motivating its exclusive use for training and validation. Future work will investigate the adaptation and external validation of the framework on datasets such as eICU-CRD [19] to assess its generalizability across different clinical settings.

A key future objective will be to transition the prototype from a standalone research system into a clinically integrated decision-support platform, followed by evaluation through simulated clinical workflows involving clinicians. These user studies will measure the system's impact on triage efficiency, diagnostic consistency, and clinician workload, evaluating its usefulness in realistic clinical scenarios. All these efforts aim to advance the proposed modular and interpretable architecture toward real-world deployment in emergency care.

Acknowledgements. We gratefully acknowledge Georgiana Copil (Adobe Romania) for her guidance and coordination throughout the development of this work. Roxana Szomiu is supported by a grant provided by Adobe Romania.

REFERENCES

- [1] Oecd, *Health at a Glance 2025: OECD Indicators*, ser. Health at a Glance. OECD, Nov. 2025. [Online]. Available: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/health-at-a-glance-2025_a894f72e/8f9e3f98-en.pdf
- [2] European Society for Emergency Medicine (EUSEM), "European emergency medicine in numbers," <https://eusem.org>, Jul. 2020, accessed: 2026-07-06. [Online]. Available: https://eusem.org/images/European_EM_in_numbers.pdf
- [3] N. Gilboy, P. Tanabe, D. Travers, and A. M. Rosenau, "Emergency severity index (ESI): A triage tool for emergency department care, version 4," Agency for Healthcare Research and Quality (AHRQ), Publication No. 12-0014, 2012. [Online]. Available: https://sgnor.ch/fileadmin/user_upload/Dokumente/Downloads/Esi_Handbook.pdf
- [4] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," *Critical Care*, vol. 23, no. 1, p. 64, 2019. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/30795786/>
- [5] L.-H. Yao, K.-C. Leung, C.-L. Tsai, C.-H. Huang, and L.-C. Fu, "A novel deep learning-based system for triage in the emergency department using electronic medical records: Retrospective cohort study," *Journal of Medical Internet Research*, vol. 23, no. 12, p. e27008, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/34958305/>
- [6] A. Vântu, A. Vasilescu, and A. Băicoianu, "Medical emergency department triage data processing using a machine-learning solution," *Heliyon*, vol. 9, no. 8, p. e18402, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844023056104>
- [7] B. M. Porto, "Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review," *BMC Emerg. Med.*, vol. 24, no. 1, p. 219, Nov. 2024. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/39558255/>
- [8] L. Masannek, L. Schmidt, A. Seifert, T. Kölsche, N. Huntemann, R. Jansen, M. Mehsin, M. Bernhard, S. G. Meuth, L. Böhm, and M. Pawlitzki, "Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study," *J. Med. Internet Res.*, vol. 26, p. e53297, Jun. 2024. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/38875696/>

- [9] G. B. Haim, M. Saban, Y. Barash, D. Cirulnik, A. Shaham, B. Z. Eisenman, L. Burshtein, O. Mymon, and E. Klang, "Evaluating large language model-assisted emergency triage: A comparison of acuity assessments by GPT-4 and medical experts," *J. Clin. Nurs.*, Nov. 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/jocn.17490>
- [10] K. Le, J. Chen, D. Mai, and K. D. R. Le, "An evaluation on the potential of large language models for use in trauma triage," *Emergency Care and Medicine*, vol. 1, no. 4, pp. 350–367, 2024. [Online]. Available: <https://www.mdpi.com/2813-7914/1/4/35>
- [11] L. Cui, M. Wang, X. Ye, W. Wei, and X. Zhao, "Diagnostic accuracy of large language models for emergency department triage: a systematic review and meta-analysis," *BMC Emergency Medicine*, 06 2026. [Online]. Available: https://www.researchgate.net/publication/407078638_Diagnostic_accuracy_of_large_language_models_for_emergency_department_triage_a_systematic_review_and_meta-analysis
- [12] J. Lee, T. Shang, J. Y. Baik, D. A. Duong-Tran, S. Yang, L. Li, and L. Shen, "Investigating llms in clinical triage: Promising capabilities, persistent intersectional biases," *ArXiv*, vol. abs/2504.16273, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280919406>
- [13] M. Lu, B. Ho, D. Ren, and X. Wang, "TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5747–5764. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.329/>
- [14] Z. Chen, Z. Peng, X. Liang, C. Wang, P. Liang, L. Zeng, M. Ju, and Y. Yuan, "MAP: evaluation and multi-agent enhancement of large language models for inpatient pathways," *Npj Health Syst.*, vol. 3, no. 1, Jun. 2026.
- [15] M. Chen, Y. Wu, J. Ma, X. Jia, C. Gao, F. Zhao, and Y. Qiao, "Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage," *NPJ Digit. Med.*, vol. 9, no. 1, Feb. 2026.
- [16] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "MIMIC-IV-ED," *PhysioNet*, Jan. 2023, version 2.2. [Online]. Available: <https://doi.org/10.13026/5ntk-km72>
- [17] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Sci. Data*, vol. 5, no. 1, p. 180178, Sep. 2018. [Online]. Available: <https://www.nature.com/articles/sdata2018178>
- [18] Y. Zhang, H. Yu, L. Hu, H. Jin, and H. Zhang, "General modular harness for llm agents in multi-turn gaming environments," 2025. [Online]. Available: <https://arxiv.org/abs/2507.11633>
- [19] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Sci. Data*, vol. 5, no. 1, p. 180178, Sep. 2018.