Bridging Trust and Efficiency: Benchmark Dataset for Collaborative AI-Human Triage for Emergency Room Decision-Making

Anonymous ACL submission

Abstract

AI offers the important potential to enhance Emergency Room (ER) triage efficiency but a lack of trust from healthcare professionals and patients limits its adoption due to concerns over accuracy and reliability. To address this, we introduce the Collaborative Intelligence-based Clinical Triage Dataset (CICTD), a large-scale benchmark containing patient requests and ER doctor annotations for triage decision-making. Along with CICTD, we define key evaluation tasks, including diagnostic question generation, ESI level prediction, triage recommendations, and misdiagnosis prediction. Our approach emphasizes a human-in-the-loop framework, ensuring AI escalates uncertain cases to experts, balancing automation with trust and improving ER triage efficiency.

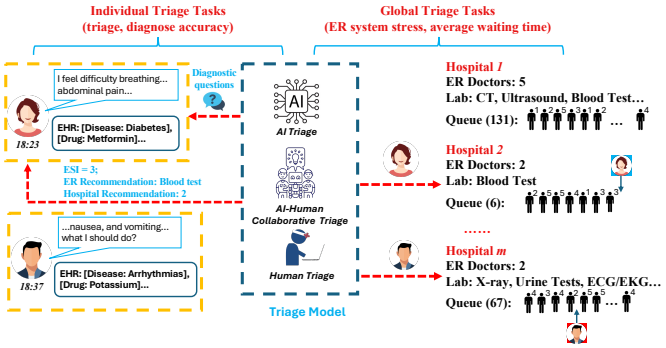

Figure 1: The CICTD dataset enables the AI-Human Collaborative Triage system to integrate patient diagnosis with ER stress management, optimizing symptom evaluation, test recommendations, and hospital assignment to reduce wait times and improve efficiency.

1 Introduction

Despite the transformative potential of AI in healthcare, generative AI techniques face significant trust challenges among health professionals and patients due to concerns over accuracy and reliability. In critical domains such as Emergency Room (ER) triage, even subtle AI errors can lead to severe consequences, undermining the adoption of AI in high-stakes decision-making. However, recent COVID-19 pandemic underscored the urgent need of efficient ER triage, as hospitals experienced overwhelming patient surges, leading to delays, misprioritized cases, and loss of lives due to resource constraints. To address these challenges, AI must be designed not as a replacement for human expertise but as a collaborative tool that supports medical professionals in high-pressure environments.

A key challenge in AI-powered ER triage is trustworthiness - AI models need to be both accurate and transparent in their decision-making. Many existing AI models, like healthcare LLMs, lack mechanisms to assess their own confidence and determine when human intervention is necessary. Additionally, there is a lack of benchmark datasets that facilitate the development and evaluation of AI models capable of collaborating with human experts. Without standardized evaluation tasks, it is difficult to measure AI performance in real-world triage scenarios, where decisions must account for both individual patients and broader hospital resource constraints. More specifically, the benchmark and evaluation tasks need to answer two critical questions: (1) How can AI be effectively integrated into healthcare, particularly ER triage? (2) How can efficient collaborative intelligence (CI) triage models be enabled to balance trust and efficiency in ER services?

In this paper, we introduce the Collaborative Intelligence-based Clinical Triage Dataset (CICTD), a large-scale benchmark dataset comprising patient requests for ER services and annotations from ER doctors on triage decision-making. Alongside CICTD, we propose a suite of evaluation tasks for triage AI models, ranging from individual tasks like diagnostic question generation, Emergency Severity Index (ESI) level prediction, and ER/triage recommendations to city-wide decisions

like triage efficiency and patient misdiagnosis prediction. These tasks not only facilitate the design of robust AI-based triage models but also support human-in-the-loop (HITL) frameworks, where AI identifies cases of low confidence and escalates decision-making to domain experts. Our work establishes a foundation for trustful and efficient ER triage systems by addressing the dual challenges of AI adoption and reliability. The CICTD dataset, coupled with the proposed evaluation framework, aims to enable the next generation of AI systems that balance automation with expert oversight, ensuring equitable and high-quality care during critical moments. Our specific contributions are:

- Introducing CICTD, a large-scale benchmark dataset with real-world patient ER requests and expert annotations for triage decisions.
- Defining key evaluation tasks along with evaluation metrics, supporting both individual and city-wide AI or CI-driven triage decisions.
- Conceptualizing a human-in-the-loop AI framework, ensuring AI models defer to experts when uncertainty is high, enhancing trust and efficiency in ER decision-making.
- Providing a foundation for the development of trustworthy AI systems in emergency medicine, bridging the gap between AI automation and expert-driven healthcare.

2 Literature Review

AI Integration in Healthcare The application of AI in healthcare has advanced rapidly with large language models (LLMs) such as GPT-4 (OpenAI, 2023), Med-PaLM (Singhal et al., 2023), and BioBERT (Lee et al., 2020), which have been employed for diagnostic support, medical documentation, and patient triage assistance. These systems have shown strong performance in processing clinical text and supporting decision-making, including AI-powered triage that aids in symptom assessment and emergency prioritization (Yi et al., 2024).

Despite these advances, several challenges impede clinical adoption. First, a lack of trustworthiness and interpretability remains a concern(Schroeder and Wood-Doughty, 2024). LLMs, while powerful, often generate responses without explicit reasoning, making it difficult for healthcare professionals to validate their outputs (Wang and Zhang, 2024). Errors, hallucinations, and biases in medical AI systems can have serious consequences (Liu et al., 2024), leading to misdiagnoses or incorrect triage decisions (Taylor et al.). Additionally, AI models struggle with uncertainty estimation (Thuy and Benoit, 2024), and they do not reliably indicate when they are unsure, making blind reliance on AI dangerous in critical medical contexts (Gao et al., 2025). Another challenge is data scarcity and annotation cost (); medical datasets require expert annotations, which are costly and time-consuming, limiting the scale and diversity of available training data.

To address these challenges, human-in-the-loop (HITL) approaches have been proposed to combine AI efficiency with human expertise in healthcare (Mosqueira-Rey et al., 2023). HITL frameworks ensure that AI models escalate uncertain cases to experts, providing decision support rather than full automation (Zhang et al., 2024). Prior work in medical AI (Lee et al., 2020) has explored expert-in-the-loop mechanisms for radiology (Fuchs et al., 2023), clinical diagnosis (Bodén et al., 2021), and medical chatbot systems (Sachdeva et al., 2024), showing improved trust and performance when AI collaborates with doctors. However, most HITL studies focus on single-instance decision-making rather than real-time adaptive collaboration in high-stakes environments (Rastogi et al., 2022), which can be problematic for ER triage adoptions. There remains a need for benchmark datasets and evaluation frameworks that support AI-human collaboration, allowing AI to learn when to defer decisions and improve medical triage efficiency (Zhang et al., 2024). Our work builds on these insights by introducing CICTD, a benchmark dataset specifically designed for AI-assisted ER triage with human-in-the-loop decision-making. We propose novel evaluation tasks, including Critical Delay Severity Assessment (CDSA), to measure and mitigate triage delays. By addressing the limitations of existing models and leveraging expert collaboration, our work provides a foundation for trustworthy, AI-assisted emergency care.

3 The CICTD dataset

Our CICTD dataset is designed to support the simulation of healthcare environments under realistic stress, which not only incorporates patient-level data but also factors such as time-of-day variations, city diversity, availability of ED phycian and different system stress settings. The dataset is divided into two components: (1) simulated ED visit records and (2) an expert-annotated subset. De-

tailed statistics are provided in Appendix 4.

3.1 Simulated ED visit records

Our patient data was sourced from the MIMIC-IV-ED v2.2 (Johnson et al., 2023) dataset and further enhanced by merging with additional records from MIMIC-IV v3.1 (Johnson et al., 2024) to capture more comprehensive ED information. The resulting dataset includes key patient attributes such as diagnosis codes, chief complaint descriptors, etc.) To accurately emulate real-world consultations, we leveraged OpenAI GPT-4 to generate detailed chief complaint description, along with a simulated round of ED physician–patient question-answer interactions. This process ensures that the synthetic text remains both clinically coherent and reflective of authentic ED data.

To build a realistic simulated patient environment, we address three key aspects. First, we stratified our patient pool based on the Emergency Severity Index (ESI) (Association, 2023), a standardized five-level triage system classifying patients by urgency and resource needs—for example, ESI level 1 denotes the highest urgency (Cairns and Kang, 2024). Using this distribution, we sampled MIMIC data to construct a dataset of 12,000 patients. The detailed ESI distribution is shown in Appendix 4.

Second, to model intra-day variations in patient arrivals, we used a time-based distribution from (Kang and Park, 2015), dividing daily visits into four periods: late night (0:00–6:00), morning (6:00–12:00, peak 1), afternoon (12:00–18:00), and evening (18:00–24:00, peak 2).

Third, to simulate varying healthcare system stress, we modeled 100 days of ED visits per city under three conditions: normal periods, flu season, and the COVID-19 pandemic. Using visit ratios of 1:3:6 from Boston Public Health data (City of Boston), we adjusted patient arrival rates to reflect increased strain, ensuring realistic seasonal and pandemic-induced variations across cities.

Emergency Department Setting We incorporate geographic diversity by selecting six cities: Boston, Houston, and Minneapolis, representing well-equipped emergency facilities, and St. Louis, Detroit, and New Orleans, reflecting high-demand, resource-constrained environments.

To define essential lab tests for ED settings, we look up ED reports (Cairns and Kang, 2024), and the MIMIC dataset, creating a standardized lab test table. Our simulation includes both high-acuity EDs and resource-limited urgent care facilities to reflect real-world diagnostic variability.

For ED physician availability, we estimate staffing needs based on annual visit volume. According to guidelines (Association, 2023), an ED physician manages 2.5 patients per hour, equating to 10,950 annual visits per physician. Using this, we approximate hospital staffing needs; for instance, an ED with 85,000 visits (Hospital, 2024) requires about eight physicians.

3.2 Expert-Annotated Dataset Subset

We generated an expert-annotated subset to address the lack of evaluation labels in simulated ED settings. A total of 300 patient records were randomly selected, ensuring coverage of all ESI Levels. These records were independently reviewed by experienced emergency physicians. Each case was annotated with key triage parameters, including maximum safe waiting time 1). These expert-labeled cases enable systematic analysis of triage decisions, resource constraints, and patient outcomes, providing a foundation for assessing the alignment between expert annotations and model-driven triage predictions, as well as evaluating resource allocation strategies.

4 Evaluation Tasks

To comprehensively evaluate both individual patient outcomes and overall system performance, we define two task levels—Individual and Global—in the ED simulation. Individual tasks assess localized triage and diagnosis, while global tasks measure large-scale resource allocation and patient prioritization.

We establish task-specific performance metrics, summarized in Table 1 and referenced in the following subsections, to ensure a structured assessment. The notation table for evaluation equations is provided in Appendix 3.

4.1 Individual Tasks

Individual tasks independently assess patient-specific predictions to improve AI-assisted triage. The structured process includes: (1) Patient Information Collection (chief complaint, history); (2) Diagnosis Question Generation (simulating physician inquiries); (3) Triage Recommendations (assigning appropriate EDs); and (4) ESI Level Prediction (estimating urgency via ESI).

By structuring tasks in this way, the evaluation captures both the accuracy of individual decision-making steps and their impact on triage efficiency.

Level	Task	Metrics	Description
Individual	Diagnosis Question Generation	$\text{NDCG}_{10}(p_i) = \frac{\sum_{j=1}^{10} r(q_{(p_i,j)}) \cdot w_j}{\sum_{j=1}^{10} r^*(q_{(p_i,j)}) \cdot w_j}$	Generates key diagnostic questions based on the chief complaint.
	Triage Recommendation	$\text{Acc}_{\text{Alloc}} = \lvert P \rvert^{-1} \sum_{p_i \in P} \delta(\hat{h}(p_i), H^*(p_i))$	Allocates patients to the appropriate ED department.
	ESI Level Prediction	$\text{ETC} = 1 - \frac{\sum_{p_i \in P} C(ESI_{p_i}, \hat{ESI}_{p_i})}{\sum_{p_i \in P} \max C(ESI_{p_i}, \cdot)})$	Assesses patient urgency based on the chief complaint.
Global	Triage Efficiency	$\text{PWWTI} = \frac{\sum_{p_i \in P} W_{\hat{ESI}_{p_i}} \cdot t_{p_i}}{\sum_{p_i \in P} W_{\hat{ESI}_{p_i}}}$	Measures the overall effectiveness of the triage system
	Real-Time Queue Stress Index	$RQSI = \min\left(1, \max\left(0, 0.6 \times \frac{T_{\text{wait}}}{x_{\text{threshold}}} + 0.4 \times \frac{\max(T_{\text{wait}} - x_{\text{threshold}}, 0)}{x_{\max} - x_{\text{threshold}}}\right)\right)$	Evaluates healthcare system load in a simulation.
	Critical Delay Severity Assessment	$\text{CSDA} = \sum_{p_i \in P} \left[W_{\hat{ESI}_{p_i}} \cdot \max\left(0, \frac{t_{p_i} - T_{\max,p_i}}{T_{\max,p_i}}\right)\right]$	Calculates the percentage of annotated patients exceeding the maximum safe waiting time

Table 1: Evaluation tasks and corresponding metrics. Most notations are self explanatory (e.g., p_i for the i'th patient) and the full list is provided in Appendix.

Diagnosis Question Generation (DSQ). Generated key diagnostic questions mimicking the inquiry process of emergency physicians during triage. The generated questions are then evaluated for their clinical relevance and essentiality in a real-world emergency department context, ensuring that they effectively support accurate and timely patient assessment.

Triage Recommendations (TR). This task aims to assess whether a model correctly assigns a patient to an emergency department (ED) that can fulfill their actual lab test requirements. The evaluation checks if the assigned hospital can meet the patient's actual lab test needs.

ESI Level Prediction. . This task assesses the model's accuracy in assigning Emergency Severity Index (ESI) levels, ranging from Level 1 (most urgent) to Level 5 (least urgent), with misclassification penalties weighted based on severity to reflect the clinical impact of incorrect triage decisions.

4.2 Global (city-level) Tasks

Unlike individual tasks that focus on optimizing patient-specific decisions, global tasks assess the overall efficiency and resource management of the emergency department network. These tasks evaluate system-wide metrics to ensure optimal patient flow and operational effectiveness across the city.

Triage Efficiency (TE). This task simulates a clinical scenario where all patients undergo triage and complete their care. Efficiency is assessed using an ESI-weighted average of waiting times, reflecting system-wide throughput and resource allocation while ensuring timely care for high-acuity cases.

Critical Delay Severity Assessment(CSDA) This task measures the proportion of annotated patients whose waiting times exceed expert-defined safety thresholds. By integrating these patients into the simulation, it evaluates the system's ability to uphold safety standards and ensure timely care.

Simulation Stress Level This task quantifies the overall burden on the emergency department(ED) system by analyzing real-time patient load, ED resource utilization and waiting queue congestion by leveraging queuing theory principles and stress-testing methodologies. We adopt expected patient waiting time as an indicator of simulation system stress, reflecting both real-time congestion and the efficiency of resource allocation.

4.3 Preliminary Experiment Results

We ran a preliminary simulation, where a Simulated Human optimizes patient outcomes by selecting hospitals and seeking direct consultations. We evaluated open-source models (Llama, Qwen) and a closed-source model (ChatGPT). As shown in Table 2, Qwen achieves the highest efficiency (TE = 123.26, CSDA = 22.22%), reducing queue stress. All models struggled with ESI prediction (Qwen2.5: 0.7433, Llama3-1: 0.7268), highlighting risks in patient emergency assessment. This underscores the challenge of balancing individual care with system-wide optimization in AI-assisted triage.

5 Conclusion

In this work, we address the critical challenge of integrating AI into ER triage by introducing the Collaborative Intelligence-based Clinical Triage Dataset (CICTD) and defining key evaluation tasks. Our benchmark supports both individual and city-level triage decision-making while emphasizing a human-in-the-loop framework to ensure AI defers to experts in uncertain cases. By bridging automation with expert oversight, our work lays the foundation for trustworthy and efficient AI-driven ER triage, ultimately improving patient outcomes and healthcare resource management.

6 Limitations

Despite the promising contributions of our work, several limitations warrant discussion:

Limited Dataset Scale Due to High Annotation Costs: The reliance on expert annotations by ER doctors constrains the dataset size because their time is both scarce and expensive. This limitation challenges the scalability of our approach. Future iterations, such as CICTD 2.0, could mitigate this issue by incorporating AI-assisted labeling combined with human verification to reduce the burden on experts.

Static Annotation of Diagnostic Questions: Diagnostic questions were annotated simultaneously, which does not fully capture the adaptive and dynamic nature of real-world triage interactions. In clinical practice, questioning evolves in response to patient feedback. Future work should aim to develop dynamic annotation methods—such as adaptive question generation—to better reflect the iterative decision-making process in emergency department triage.

Trade-offs Between Individual and Global Optimization: Our study highlights a key challenge in triage systems: balancing individual patient care with overall system efficiency. The "Simulated Human" model, which focuses solely on maximizing individual patient care, achieves excellent performance in patient-centric tasks (e.g., triage recommendations and diagnostic question generation), but this focus comes at the expense of system-wide efficiency. Conversely, global optimization techniques, as demonstrated by large models like Llama3-1 and Qwen2.5, improve overall triage efficiency while compromising accuracy in individual tasks such as ESI level prediction. This trade-off is particularly critical when minor errors in assessing high-risk patients can have severe consequences.

Challenges in Integrating Human-in-the-Loop Approaches: While a hybrid approach combining machine intelligence with human decision-making shows promise, directly applying NLP techniques like Mixture of Experts (MoE)(Shazeer et al., 2017) in emergency department triage is less straightforward. Although MoE strategies improve efficiency in NLP by activating only a subset of experts based on input tokens, ED triage demands immediate and comprehensive consideration of patient needs, limiting the applicability of such methods.

Addressing these limitations will be crucial in future research to develop a more robust and efficient emergency department triage system that balances individual patient care with global system performance.

Acknowledgments

References

Emergency Nurses Association. 2023. Emergency Severity Index Handbook, Fifth Edition. Emergency Nurses Association, Schaumburg, IL. Available online at https://www.ena.org.

Anna CS Bodén, Jesper Molin, Stina Garvin, Rebecca A West, Claes Lundström, and Darren Treanor. 2021. The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice. Histopathology, 79(2):210–218.

Christopher Cairns and Kai Kang. 2024. National hospital ambulatory medical care survey: 2022 emergency department summary tables.

City of Boston. Respiratory illness dashboard. https://www.boston.gov/departments/public-health-commission/respiratory-illness-dashboard. Accessed: 2025-02-16.

Moritz Fuchs, Camila Gonzalez, Yannik Frisch, Paul Hahn, Philipp Matthies, Maximilian Gruening, Daniel Pinto Dos Santos, Thomas Dratsch, Moon Kim, Felix Nensa, et al. 2023. Closing the loop for ai-ready radiology. In RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren. Georg Thieme Verlag KG.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy Miller, Danielle S Bitterman, Guanhua Chen, Anoop Mayampurath, Matthew M Churpek, and Majid Afshar. 2025. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA open, 8(1):ooae154.

Barnes-Jewish Hospital. 2024. Barnes-jewish hospital emergency department overview. Accessed: February 16, 2025.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023. Mimic-iv-ed (version 2.2). PhysioNet. Published: Jan. 5, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Benjamin Gow, Brian Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. Mimic-iv (version 3.1). PhysioNet.

Seung Woo Kang and Hyun Soo Park. 2015. Emergency department visit volume variability. Clinical and experimental emergency medicine, 2(3):150.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon 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.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. 2024. Exploring and evaluating hallucinations in llm-powered code generation. arXiv preprint arXiv:2404.00971.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56(4):3005–3054.

OpenAI. 2023. Gpt-4 technical report. ArXiv:2303.08774.

Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2022. A unifying framework for combining complementary strengths of humans and ml toward better predictive decision-making. arXiv preprint arXiv:2204.10806.

Bhuvan Sachdeva, Pragnya Ramjee, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. Learnings from a large-scale deployment of an llm-powered expert-in-the-loop healthcare chatbot. arXiv preprint arXiv:2409.10354.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust llm judgments? reliability of llm-as-a-judge. arXiv preprint arXiv:2412.12509.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. Preprint, arXiv:1701.06538.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

R Andrew Taylor, Rohit B Sangal, Moira E Smith, Adrian D Haimovich, Adam Rodman, Mark S Iscoe, Suresh K Pavuluri, Christian Rose, Alexander T Janke, Donald S Wright, et al. Leveraging artificial intelligence to reduce diagnostic errors in emergency medicine: Challenges, opportunities, and future directions. Academic Emergency Medicine.

Arthur Thuy and Dries F. Benoit. 2024. Explainability through uncertainty: Trustworthy decision-making with neural networks. European Journal of Operational Research, 317(2):330–340.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. Artificial Intelligence Review, 57(11):299.

Nayeon Yi, Dain Baik, and Gumhee Baek. 2024. The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies. Journal of Nursing Scholarship.

Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, page 1–18. ACM.

A Appendix

A.1 Simulation Assumption and Method

For this study, we adopt a simplified model to assess the feasibility of our triage chatbot in an emergency setting. Given the complexity of real-world emergency department operations, our model is designed to capture the essential elements of the triage and treatment process while maintaining clarity and efficiency in simulation.

First, we assume that the time required for triage is consistent across all patients, regardless of the number of individuals waiting to be assessed. While, in reality, triage time may vary depending on patient volume and the severity of each case, our model applies an average triage time that reflects typical emergency department conditions.

Second, our model features a single queue for the doctor, where patients are seen in the order they arrive after triage. While there is technically a separate queue for the triage nurse in real-world settings, we simplify this aspect by assuming all patients experience the same average triage time before entering the doctor's queue.

Third, patient wait times for a doctor are determined by two key factors: (1) position in the queue, which is based on arrival order after triage; and (2) doctor availability, as each doctor can only treat one patient at a time, though they may manage multiple cases simultaneously up to a predefined capacity.

By focusing on these fundamental assumptions, our model provides a structured yet adaptable foundation for evaluating the role of triage automation in emergency care. Future iterations may incorporate additional complexities such as dynamic triage wait times, real-time re-prioritization based on patient condition, and variability in triage nurse efficiency.

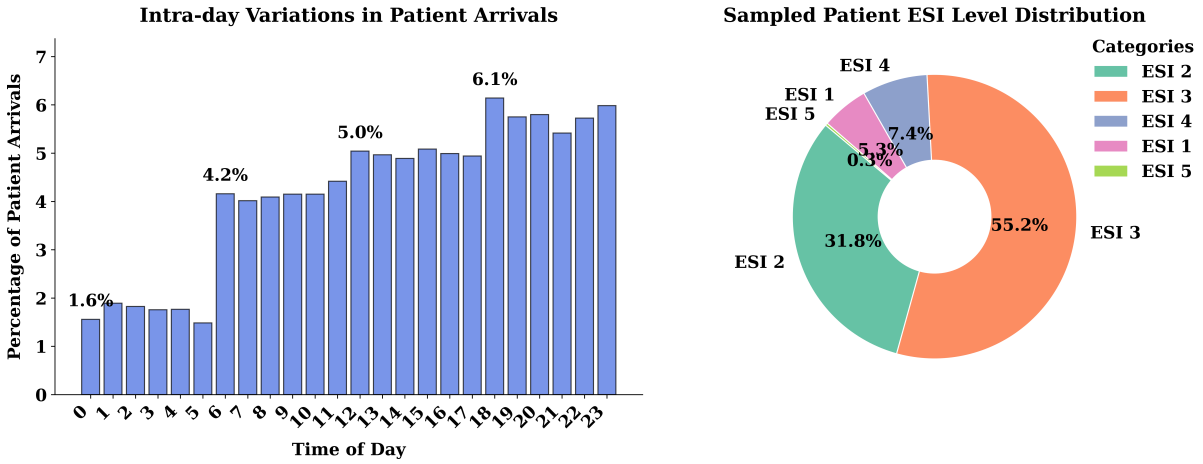

Figure 2: CICTD Dataset Patient Distribution in intra-day arrival and ESI level .

Model	Params	Individual			Global		
		DQG	TR	ESI Prediction	TE	CSDA	OSR
Simulated Human	—	1.0	0.8840	1.0	1024.40	50%	0.5611
Qwen2.5	7B	0.9366	0.9180	0.7433	123.26	22.22%	0.0945
Llama3-1	8B	0.9104	0.8528	0.7268	638.31	44.44%	0.4740
ChatGPT	—	0.9291	0.9875	0.8182	832.77	38.80%	0.4680

Table 2: Performance comparison of different models

Laboratory and Diagnostic Tests

Core Laboratory Tests

- Complete Blood Count (CBC)
- Basic Metabolic Panel (BMP)
- Liver Function Tests (LFTs)
- Coagulation Studies
- Urinalysis (UA)
- Type and Screen
- Electrolytes (Ionized Calcium, Magnesium, Phosphate)
- Renal Function Tests (BUN, Creatinine Clearance)
- Inflammatory Markers (CRP, ESR, Procalcitonin)

Advanced Tests

Point-of-Care Testing (POCT)

- Fingerstick glucose
- Lactate level
- Venous/Arterial Blood Gas (VBG/ABG)
- PT/INR
- Urine hCG (for all reproductive-age females)
- Rapid influenza/COVID-19 antigen tests

Critical Time-Sensitive Tests

- Serial lactate
- Troponin-I/T
- D-dimer
- Lipase
- BNP (Brain Natriuretic Peptide)
- CK-MB (Creatine Kinase MB)

Toxicology & Drug Screens

- Salicylate (Aspirin Level)
- Acetaminophen Level
- Carboxyhemoglobin
- Methemoglobin
- Opiates, Cocaine, Benzodiazepines
- Ethanol Level

Infectious Disease Tests

- HIV Screening (HIV Ag/Ab)
- Hepatitis Panel (HBsAg, Anti-HCV, etc.)
- COVID-19, Influenza, RSV
- Syphilis (RPR, Treponemal Ab)

Imaging Studies

Basic Imaging

- X-ray
- EKG (ECG)

Advanced Imaging

- Ultrasound
- CT scan
- MRI

Notation	Description
S	Medical simulation scenario
T	Total number of days in the simulation S
$P = \{p_1, p_2, ..., p_n\}$	Set of patients
$H = \{h_1, h_2, ..., h_m\}$	Set of hospitals h
$L\hat{a}b_{h_j}, L\hat{a}b_{h_{ij}}$	Set of lab tests supported by hospital h_j
Doc_{h_j}	Number of available doctors in hospital h_j
$Queue_{h_j}$	Queue length in hospital h_j
C_{p_i}	Chief complaint description of patient p_i
M_{p_i}	Past medical history of patient p_i
t_{p_i}	Time when patient p_i enters the scenario
$E\bar{S}I_{p_i}, E\hat{S}I_{p_i}$	Ground truth and predicted ESI level of patient p_i, where $E\bar{S}I_{p_i} \in \{1, 2, 3, 4, 5\}$
$Q_{p_i} = \{q_{(p_i,1)}, ..., q_{(p_i,10)}\}$	10 Diagnosis Questions generated for patient p_i, based on C_{p_i} and M_{p_i}
$\hat{h}_{p_i}$	Hospital ID predicted by the model as the most suitable allocation for patient p_i
$\bar{H}_{p_i}$	Set of hospitals that support the ground truth lab test recommendation for patient p_i
t_{p_i}	The actual waiting time for patient p_i in simulation S
$W_{E\bar{S}I_{p_i}}$	Predefined urgency weight assigned based on $E\bar{S}I_{p_i}$
$T_{\max,p_i}$	The max waiting time for patient p_i that is annotated by the ED doctor.

Table 3: Notation Table for Medical Simulation Scenario

Category	Statistic	Value / Distribution
Overall Scale	Total ED Visits	12000
	Unique Patients	11407
	Expert-Annotated Cases	300
	% ED Visits with Discharge Report	42.58%
	Simulation Time Span	50 Days. (1/3 Normal, 1/3 flu, 1/3 Covid periods)
	# of Cities	6 (Boston, Houston, Minneapolis, St. Louis, Detroit, New Orleans)
	# of Lab Test Covered	36 (CBC, BMP, etc)
ESI Distribution (# of Patients)	ESI 1	640
	ESI 2	3817
	ESI 3	6618
	ESI 4	892
	ESI 5	33
Hourly Patient Arrival Distribution	Late Night (0:00–6:00)	Mean: 34.91, Std: 21.15
	Morning (6:00–12:00)	Mean: 85.714, Std: 51.71
	Afternoon (12:00–18:00)	Mean: 102.71, Std: 62.51
	Evening (18:00–24:00)	Mean: 119.51, Std: 72.93
ED Outcomes	% of Admitted	31.45%
	% of Went Home	62.56%

Table 4: Emergency Department Statistics