Privacy-First Triage Classification with Open-Weight LLMs: A Chain-of-Thought Distillation Approach

Zeyuan Zhao
Montgomery Blair High School
Silver Spring, USA
azhaodev@gmail.com

Yexiao He, Ang Li
Department of Electrical and Computer Engineering
University of Maryland College Park
College Park, USA
{yexiaohe,angliece}@umd.edu

***Abstract*—Hospital triage is a critical, but often error-prone step in emergency department (ED) care. Misclassifications can delay necessary care and decrease resource efficiency. Recent studies suggest large language models (LLMs) are promising in ED decision-making. However, most prior work has focused on closed-weight models and baseline evaluations without exploring additional strategies. This paper investigates adapting a lightweight, open-weight model for ED triage. We fine-tune gpt-oss-20b using Low-Rank Adaptation on publicly available records from the MIMIC-IV dataset and present a student-teacher chain-of-thought distillation method to improve performance. Various Low-Rank Adaptation configurations are evaluated in an ablation study and are compared across several metrics. Our best model improves over the base gpt-oss-20b by almost 18% and compares to the real-world nurse triage accuracy. The results suggest that smaller, open-weight models can feasibly support ED triage. Such models can be deployed locally, preserving privacy and reducing reliance on proprietary models. These findings provide evidence that open-weight large language models optimized with techniques such as model distillation are able to perform effectively on clinical tasks.**

***Index Terms*—Large language models, Machine learning**

I. Introduction

Patients seeking urgent medical care arrive at the Emergency Departments (EDs) of hospitals and are classified in a process called triage. ED nurses take in data points, including chief complaint, pain level, a brief history, and important vitals, then categorize patients based on acuity, ensuring efficient resource allocation and prioritization of severe cases [6].

A widely used triage algorithm in the United States is the Emergency Severity Index (ESI), ranging from 1 (highest acuity) to 5 (lowest acuity). The ESI classifies patients firstly on urgency/risk of deterioration, then on resources required. If a patient requires immediate intervention, a score of 1 is assigned. Less immediate but still urgent cases are assigned a score of 2. Patients who are determined to be stable are then classified levels 3 through 5 based on resources required [6].

Triage is a highly complex task, relying on ED nurses' judgement and interpretation of many data points. A previous study placed real-world triage accuracy to be around 59% [6]. Many factors, including overcrowding or inexperience, can lead to mistriage. Undertriage, the erroneous assignment of a low acuity, can lead to essential care being delayed. Although overtriage errs on the side of caution, it causes inefficient resource allocation [2].

Large language models (LLMs) are able to effectively reason over complicated, free-form text. Recent work has explored applying them to multiple clinical tasks [11]. Gaber *et al.*, in particular, ran several Claude models on 3 tasks: triage assignment, specialty referral, and diagnosis prediction [12]. While the study demonstrated that LLMs are quite feasible in clinical tasks, the paper primarily served as a baseline benchmark and did not investigate any new techniques, other than Retrieval Augmented Generation (RAG) [13].

This paper will focus on the triage task and explore techniques for increasing prediction accuracy using open-weight models beyond improved prompting. A smaller, open-weight model was purposely chosen to demonstrate feasibility for use in a real-life ED setting. Patient privacy regulations sometimes restrict the transfer of sensitive medical records over the internet [16]. Many consumer machines, which hospitals may use, are capable of running smaller models locally. Reducing reliance on proprietary systems not only improves patient privacy but may also reduce costs [15].

We construct a high-quality dataset with clinical information and ground truths, used both for training and testing. We then propose and implement a model distillation process while comparing various fine-tuning hyperparameter configurations. Finally, we benchmark the trained models against the baseline models.

II. Related Work

A. Medical Applications of LLMs

Previous works have used LLMs for a variety of medical tasks such as exam question answering, clinical note summarization, and medical education [16]. In terms of prediction-related tasks, LLMs have been used to process medical sensor recordings [7], diagnose patients [14], triage patients [8], and more [11].

Privacy and ethical concerns must be considered when employing LLMs to process sensitive patient records. Open-weight models, which can comply with regulations more easily

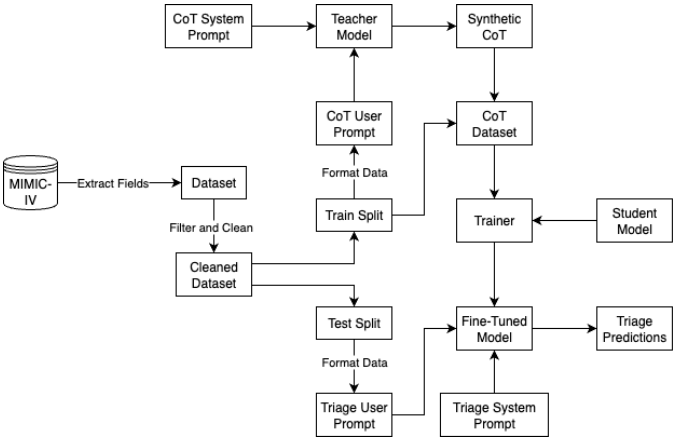

Fig. 1. Model Distillation Pipeline

due to being self-hosted, have shown limitations in terms of performance, motivating research in training and inference strategies [16].

B. Chain-of-Thought Distillation

Smaller models, which may struggle to match the performance of more complex models, have shown improvement on medical tasks through chain-of-thought (CoT) generation. [16]. Additionally, LLMs can also be fine-tuned on specific medical tasks, significantly increasing performance [10]. Recent work has shown that further improvement can be made via distilling example CoTs. A larger "teacher" model is prompted to generate example CoTs, which are used to fine-tune a lightweight "student" model, allowing it to learn complex reasoning [4].

C. Triage Prediction

Prior work on triage prediction has relied on traditional machine learning models [3]. More recently, LLMs have been explored for ED triage [9]. Gaber *et al.* evaluates Claude on ESI classification [12]. Retrieval augmented generation was also employed for the triage task [17].

III. Methods

The goal of the study was to develop an LLM framework that would mimic the restrictions of a real ED setting to explore the feasibility for deployment. We chose several models that were both open-weight and small to work with. To benchmark and improve the models, we prepared a dataset derived from real medical records. We generated synthetic CoT examples for model distillation [4]. We then test various Low-Rank Adaptation (LoRA) [1] configurations and compare them against the base model.

We define patient records as structured feature vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, where i is the ith patient and m represents the mth input feature (e.g. heart rate, chief complaint). The ground truth triage level is represented as a label $y_i \in \{1, 2, 3, 4, 5\}$. A prediction model is a function f_θ with parameters θ, which maps x_i to a prediction $\hat{y}_i = f_\theta(x_i)$.

We can define a dataset $S = \{(x_i, y_i)\}_{i=1}^N$, where S_{test} and S_{train} represent the testing and training set, respectively.

Let T be the teacher model, which generates a CoT $r_i = T(x_i, y_i)$ for patient i. We can then construct a CoT training dataset $S_{CoT} = \{(x_i, r_i, y_i)\}_{i=1}^N$. The fine-tuning process can be represented as $\Delta\theta \leftarrow \text{Train}(f_\theta, S_{CoT})$ and $\theta^* = \theta + \Delta\theta$, where $\Delta\theta$ represents the learned LoRA adapters. A high-level summary of the pipeline is as in equation (1):

$$S_{train} \xrightarrow{T} S_{CoT} \xrightarrow{\text{Train}} f_{\theta^*} \tag{1}$$

A. Model

For this paper, we chose the model with 2 criteria: (1) small size, and (2) open-weight. OpenAI's gpt-oss-20b is an open-weight Mixture-of-Experts (MoE) model, and has shown promise in benchmarks. Its MoE architecture and MXFP4 quantization allow for fast inference and a memory requirement of only 16GB, respectively [15].

B. Dataset

Our dataset was created using records found in the `MIMIC-IV 3.1` dataset [5], which includes data on over 200,000 patients admitted to the ED at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. We used the `Hosp`, `ED`, and `Note` modules, and extracted fields including `gender`, `arrival_transport`, `text`, etc. The `acuity` field was used as the ground truth, and the others as input data.

Patients with any empty fields were discarded. The text field contained each patient's entire discharge record, which included future info (e.g., diagnosis, labs) that would give the model unfair hindsight. We extracted only the history of present illness for the triage task and used RegEx rules to remove irrelevant information and potential hindsight.

The number of viable patients after filtering was 22,948. We used 12,000 patients for S_{train} (for which we generated CoT), and 2,000 patients for S_{test}.

C. Chain-of-Thought Distillation

gpt-oss-20b generates a CoT prior to outputting an answer. In order to finetune the model, we provide the trainer with CoT examples. Therefore, we adopt a teacher-student model, where a more sophisticated teacher model T uses S_{train} to generate synthetic CoT dataset S_{CoT}, which is then used to train our student base prediction model f_θ.

We chose GPT-5 for T. We provide T specific instructions in the system prompt to interpret all relevant patient information in the context of the ESI system, allowing the student model to learn the reasoning patterns. See *Appendix A* for the full prompts and templates.

D. Low-Rank Adaptation

In order to train the model efficiently, we use LoRA [1], a parameter-efficient fine-tuning method. Rather than updating all parameters, LoRA works by freezing the pre-trained weights and injecting lightweight, trainable adapters in certain

layers. This method reduces the number of trainable parameters by several magnitudes, requiring less computational resources. We also applied LoRA instead of full fine-tuning to reduce the chances of overfitting and catastrophic forgetting.

Let $W_l \in \mathbb{R}^{d_{out} \times d_{in}}$ be the weight matrix for a given layer l. If r represents the rank of the matrix decomposition and α is a scaling factor, an update is then

$$\Delta W_l = \frac{\alpha}{r} \cdot B_l A_l \tag{2}$$

, where $A_l \in \mathbb{R}^{r \times d_{in}}$ and $B_l \in \mathbb{R}^{d_{out} \times r}$. A higher r allows for more sophisticated learning at the cost of a larger training footprint. The α value controls the strength of the adaptation.

The full set of LoRA parameters can be represented with $\Delta\theta = \{\Delta W_l\}_{l \in \mathcal{L}}$, where $\mathcal{L}$ denotes the set of adapted layers. During inference, the weights may be merged by $W_l' = W_l + \Delta W_l$, allowing for lower latency.

For this paper, we applied LoRA adapters to the transformer's attention projections. We used LoRA and S_{CoT} to adapt the base model f_θ to the triage task, obtaining f_{θ^*}.

IV. Experiments

We conducted experiments by evaluating several model configurations on S_{test}. Firstly, we evaluated the base gpt-oss-20b model to serve as a baseline. The student model was then fine-tuned with several hyperparameter configurations for an ablation study on S_{CoT} using LoRA as described in *III. Methods*. We use a variety of widely used metrics to compare the effectiveness of the different fine-tuning configurations.

A. Setup and Configuration

The experiments were conducted on RTX 6000 Ada GPUs, each with 48GB of memory. We used 4 GPUs for the fine-tuning phase, and 1 GPU for evaluation. The model, gpt-oss-20b, was quantized using MXFP4 during inference.

We define the control fine-tuned model with the configuration in *Table I*. For the ablation study, we varied the value for rank, alpha, and learning rate. Training loss was logged at every step, while evaluation on the validation set was performed every 25 steps to monitor model generalization. Models were fine-tuned up to 1,000 steps. We trained each configuration only once due to computational limitations.

TABLE I
HYPERPARAMETER CONFIGURATION FOR THE CONTROL FINE-TUNED MODEL.

Hyperparameter	**Value**
Rank (r)	128
Alpha (α)	256 ($2r$)
Learning rate	2E-04
Dropout	0.05
Per-device batch size	2
Gradient accumulation steps	12
Warmup ratio	0.03
Weight decay	0.01
Scheduler	Cosine (min 0.1)
Optimizer	AdamW

B. Metrics

Exact-match Accuracy: A prediction $\hat{y}$ is considered correct if it exactly matches ground truth y.

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}, \tag{3}$$

Over-one Accuracy: $\hat{y}$ is considered correct if it overtriages by at most 1. This captures near-miss predictions, erring on the side of caution.

$$\text{Acc+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{0 \leq \hat{y}_i - y_i \leq 1\}. \tag{4}$$

Under-triage Rate: The proportion of cases where $\hat{y}$ is higher than y (i.e., potentially risky underestimation):

$$\text{UTR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i > y_i\}. \tag{5}$$

Over-triage Rate: The proportion of cases where $\hat{y}$ is lower than y (i.e., overly cautious prediction):

$$\text{OTR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i < y_i\}. \tag{6}$$

C. Results

We conducted an ablation study on several hyperparameter configurations, which vary from the control fine-tuned model. Afterwards, we combined the best-performing hyperparameter settings and tested the model. The specific configurations and their performance on the metrics throughout several checkpoints can be found in *Table II*.

For the final checkpoint, the highest accuracy configuration scored an accuracy of **62.53%**, which is **17.97%** higher than the base accuracy of 44.56%. Compared to the real-world accuracy rate of 59% [6], our best model outperforms by **3.68%**. The $r = 256$ model improved 2.11% over the control fine-tuned model, and also outperformed the other configurations on most of the metrics. The fine-tuned models have lower UTR, while gpt-oss-20b has the lowest OTR.

V. Discussion

Our results demonstrate that CoT distillation with LoRA substantially improves the performance of gpt-oss-20b for ED triage classification. We observed the highest gains with increased rank ($r = 256$). The improvement in Acc and Acc+1 indicate the model's ability to capture complex reasoning patterns. Reductions in UTR and increases in OTR all over the board indicate the model's preference for caution upon fine-tuning.

The ablation study results show that both the learning rate and the rank-alpha combination were influential. Higher rank generally resulted in superior performance, though the difference became less pronounced as training progressed. This suggests that overfitting was not an issue. Learning rate

TABLE II
METRICS FOR ABLATION STUDY ON SEVERAL FINE-TUNING CONFIGURATIONS AT SELECTED CHECKPOINTS.

Fine-tune Configuration	Step 250				Step 500				Final/Step 1000			
	Acc	Acc+1	UTR	OTR	Acc	Acc+1	UTR	OTR	Acc	Acc+1	UTR	OTR
Base gpt-oss-20b									44.56	56.36	43.24	**12.20**
Control $r = 128$ $\alpha = 256$ LR = 2E-04	58.37	74.50	25.00	**16.63**	60.67	77.56	22.04	**17.28**	60.42	78.26	21.24	18.34
Ablation $r = 256$ $\alpha = 512$ LR = 2E-04	**60.22**	80.51	18.84	20.94	59.72	81.06	18.19	22.09	62.53	**79.16**	**20.24**	17.23
Ablation $r = 64$ $\alpha = 128$ LR = 2E-04	56.99	77.94	21.50	21.50	60.73	77.98	21.56	17.70	60.71	78.78	20.52	18.77
Ablation $r = 128$ $\alpha = 256$ LR = 4E-04	57.31	78.76	20.69	22.00	**61.64**	**81.63**	**17.66**	20.70	**62.68**	78.84	20.61	16.71
Ablation $r = 128$ $\alpha = 256$ LR = 1E-04	56.07	74.20	25.15	18.78	59.37	78.00	21.50	19.14	61.07	78.45	20.89	18.03
Ablation $r = 128$ $\alpha = 256$ LR = 5E-05	56.35	77.82	21.47	22.18	56.34	75.72	23.67	19.99	57.29	78.24	21.25	21.45
Ablation $r = 128$ $\alpha = 128$ LR = 2E-04	57.79	78.39	21.11	21.11	59.62	78.70	20.80	19.59	60.81	78.99	20.25	18.94
Combined $r = 256$ $\alpha = 512$ LR = 4E-04	56.56	**84.56**	**14.68**	28.76	59.27	80.14	19.25	21.48	60.13	77.05	22.59	17.27

significantly affected the final Acc, with the highest, 4E-04, performing the best. Setting the learning rate too low led to weaker convergence and higher error rates. Training with $\alpha = r$ did not significantly affect performance. Although both the $r = 256$ and the $LR =$ 4E-04 configurations performed the best in their respective categories, the combined configuration model notably did not improve from the base model.

Several limitations may constrain the interpretation of these results. Each configuration was only trained and tested once, limiting statistical confidence. The results of this paper may not fully cover all real-world triage scenarios due to the dataset construction method, and only using the MIMIC-IV dataset. Triage labels in MIMIC reflect human-assigned assessments, which are noisy and not guaranteed to be clinically correct. The real-world accuracy of 59% should be taken into account when interpreting results.

VI. CONCLUSION

In this paper, we explored the feasability of deploying cost-efficient large language models for ED triage classification using teacher-student model distillation. We generated synthetic CoTs from the teacher model, GPT-5, to train the smaller, open-weight gpt-oss-20b using LoRA. The ablation study showed that the best configuration improved reasoning robustness and increased the accuracy by nearly 18%. This work highlights the promise of lightweight, open models in supporting real-world triage workflows, taking into consideration privacy concerns and computational restrictions. Future research should test on more diverse patient cohorts and explore extending to related tasks such as specialty referral or diagnosis prediction.

REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, *et al.*, *LoRA: Low-Rank Adaptation of Large Language Models*, arXiv:2106.09685 [cs], Oct. 2021. DOI: 10.48550/arXiv.2106.09685.

[2] T. Levis-Elmelech, D. Schwartz, and Y. Bitan, "The effect of emergency department nurse experience on triage decision making," *Human Factors in Healthcare*, vol. 2, p. 100 015, Dec. 2022. DOI: 10.1016/j.hfh.2022.100015.

[3] R. Sánchez-Salmerón, J. L. Gómez-Urquiza, L. Albendín-García, *et al.*, "Machine learning methods applied to triage in emergency services: A systematic review," *International Emergency Nursing*, vol. 60, p. 101 109, Jan. 2022. DOI: 10.1016/j.ienj.2021.101109.

[4] N. Ho, L. Schmid, and S.-Y. Yun, *Large Language Models Are Reasoning Teachers*, arXiv:2212.10071 [cs], Jun. 2023. DOI: 10.48550/arXiv.2212.10071.

[5] A. E. W. Johnson, L. Bulgarelli, L. Shen, *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," en, *Scientific Data*, vol. 10, no. 1, p. 1, Jan. 2023, Publisher: Nature Publishing Group. DOI: 10.1038/s41597-022-01899-x.

[6] Lisa Wolf, Katrina Ceci, Danielle McCallum, Deena Brecher, Deb Jeffries, and Rebecca McNair, *Emergency Severity Index Handbook Fifth Edition*, 2023.

[7] X. Liu, D. McDuff, G. Kovacs, *et al.*, *Large Language Models are Few-Shot Health Learners*, arXiv:2305.15525 [cs], May 2023. DOI: 10.48550/arXiv.2305.15525.

[8] L. Masanneck, L. Schmidt, A. Seifert, *et al.*, "Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study," EN, *Journal of Medical Internet Research*, vol. 26, no. 1, e53297, Jun. 2024, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. DOI: 10.2196/53297.

[9] C. Y. K. Williams, T. Zack, B. Y. Miao, *et al.*, "Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department," *JAMA Network Open*, vol. 7, no. 5, e248895, May 2024. DOI: 10.1001/jamanetworkopen.2024.8895.

[10] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning Large Language Models for Specialized Use Cases," English, *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, Mar. 2025, Publisher: Elsevier. DOI: 10.1016/j.mcpdig.2024.11.005.

[11] S. Bedi, Y. Liu, L. Orr-Ewing, *et al.*, "Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review," *JAMA*, vol. 333, no. 4, pp. 319–328, Jan. 2025. DOI: 10.1001/jama.2024.21700.

[12] F. Gaber, M. Shaik, F. Allega, *et al.*, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," en, *npj Digital Medicine*, vol. 8, no. 1, p. 263, May 2025, Publisher: Nature Publishing Group. DOI: 10.1038/s41746-025-01684-1.

[13] O. K. Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *Digital Health*, vol. 11, p. 20 552 076 251 337 177, Apr. 2025. DOI: 10.1177/20552076251337177.

[14] G. K. Gupta, P. Pande, N. Acharya, A. K. Singh, and S. Niroula, *LLMs in Disease Diagnosis: A Comparative Study of DeepSeek-R1 and O3 Mini Across Chronic Health Conditions*, arXiv:2503.10486 [cs], Jun. 2025. DOI: 10.48550/arXiv.2503.10486.

[15] OpenAI, S. Agarwal, L. Ahmad, *et al.*, *Gpt-oss-120b & gpt-oss-20b Model Card*, arXiv:2508.10925 [cs], Aug. 2025. DOI: 10.48550/arXiv.2508.10925.

[16] J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, "A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration," en, *Healthcare*, vol. 13, no. 6, p. 603, Jan. 2025, Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/healthcare13060603.

[17] M. Yazaki, S. Maki, T. Furuya, *et al.*, "Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model," *Prehospital Emergency Care*, vol. 29, no. 3, pp. 203–209, Apr. 2025, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10903127.2024.2374400. DOI: 10.1080/10903127.2024.2374400.

APPENDIX A
PROMPTS

A. Triage Prediction System Prompt

You are a nurse with expertise in determining triage level using patient info.
Task: Assign a triage level using the Emergency Severity Index (ESI), which ranges from 1 (highest acuity) to 5 (lowest acuity):
Instructions:
- Include and interpret all relevant info (vitals, complaints, history, etc.) per ESI guidelines.
- Explain how abnormal each piece of info is and how it affects or does not affect acuity.
- Write in 2-3 paragraphs describing each relevant finding, its interpretation, and its effect on ESI. Discuss resources needed and why each ESI is met or not met.
- Ignore all future tests, labs, imaging, etc. that are given after acuity assessment that may accidentally be present from data leakage.
- Identify the most appropriate triage level based on the provided information. When in doubt, choose the higher acuity to prioritize safety.
- Respond only with one <acuity> tag with an ESI level ranging from 1 to 5.
ESI levels:
1: Immediate lifesaving intervention.
2: High-risk situation or severe distress (e.g., confused, lethargic, disoriented, severe pain/distress).
3: Many interventions required. Vitals are non-urgent.
4: One intervention required (e.g., lab test, imaging, or EKG).
5: No interventions beyond exam.
Example response:
<acuity>[integer 1-5]</acuity>

B. CoT Generation System Prompt

You are a medical expert in hospital triage. You will be given patient data and the actual assigned ESI level.
Task: Generate a realistic rationale/reasoning leading to the correct ESI level.
Instructions:
- Include and interpret all relevant info (vitals, complaints, history, etc.) per ESI guidelines.
- Explicitly state how abnormal each piece of info is. Explain how each piece of info affects or does not affect acuity.
- Discuss only current findings. Do not list hypothetical future risks.
- Write in 2-3 paragraphs describing each relevant finding, its interpretation, and its effect on ESI. Discuss resources needed and why each ESI is met or not met.
- Reason fully and clearly as a triage nurse, explaining the decision in complete sentences.
- Ignore info that is mistakenly included or should be unavailable during triage (imaging, labs, family history, etc.).
- Conclude: Final ESI level: [level]
ESI levels:
1: Immediate lifesaving intervention.
2: High-risk situation or severe distress (e.g., confused, lethargic, disoriented, severe pain/distress).
3: Many interventions required. Vitals are non-urgent.
4: One intervention required (e.g., lab test, imaging, or EKG).
5: No interventions beyond exam.