Privacy-First Triage Classification with Open-Weight LLMs

A Chain-of-Thought Distillation Approach

Zeyuan Zhao
Montgomery Blair High School

Yexiao He, Ang Li
University of Maryland

1. BACKGROUND

- **Triage:** nurses sort patients upon entering the hospital
  - Goal: prioritize patients to ensure efficiency
  - Information: vitals, history of present illness, *etc.*
  - Score: 1 (most) to 5 (least priority)
  - Real-world accuracy reported ~59%, according to ESI Handbook
- **Large language models (LLMs)** can assist with triage
  - Most systems are proprietary and closed-weight
  - Invasive to privacy, must send over internet
  - Inaccessible in remote regions, expensive

Our goal is to create an **accurate** triage prediction system that handles **real-world** cases while deployable **locally** at **low cost**. This ensures **patient privacy** is protected and **underserved areas** have access.

STATISTICS

62.68%
vs 59% human accuracy

+18.02%
vs Base Model

<1 min
Inference Time

16 GB
RAM Required

IMPACT

✓ Preservation of privacy
- No **sensitive data** is transmitted over the internet, complies with regulations

✓ Accessible
- Free, open-weight model ensures **minimal cost barrier**, reducing health disparities

✓ Remote areas
- Local system does not require **internet access**

2. METHODS: MODELS

Student Model
Criteria
- Open-weight
- Low compute

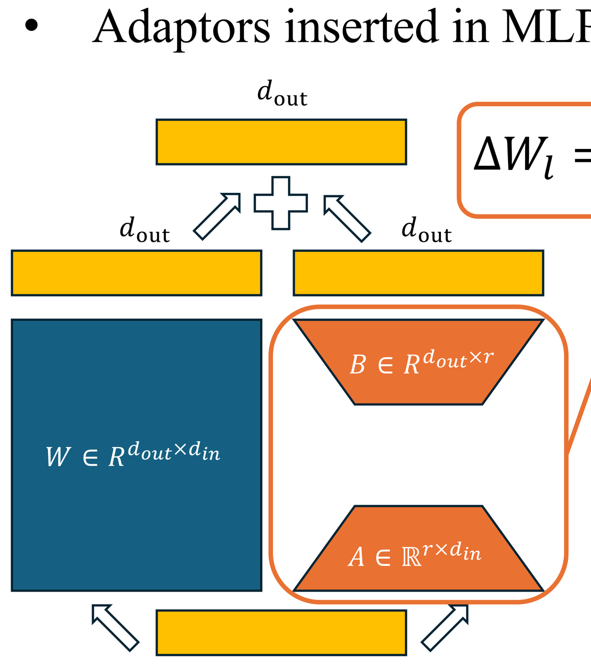
OpenAI gpt-oss-20b
- HealthBench: 42.5%
- Chain-of-Thought (CoT): better medical reasoning performance
- MXFP4 quantization: requires only 16 GB RAM
- Mixture-of-Experts: fast inference, important in hospitals

Teacher Model
Criteria
- High task performance

OpenAI GPT-5
- HealthBench: 67.2%

3. METHODS: LOW-RANK ADAPTATION

We employ low-rank adaptation (LoRA) for fine-tuning.
- Original weights are frozen
- Adaptors inserted in MLP & attention layers



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

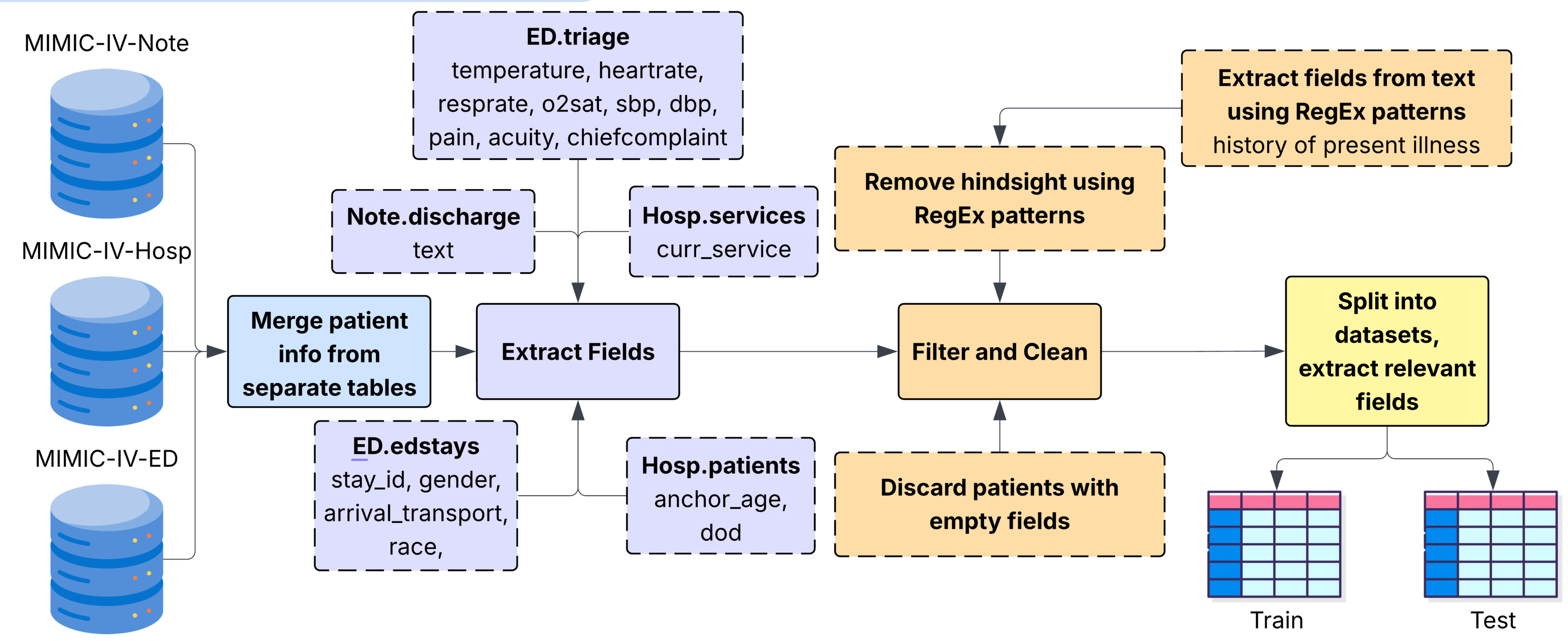
α: scaling factor of adapters
r: rank of matrix decomposition

4. METHODS: DATASET CREATION

Train on real patient records from MIMIC-IV to **assess viability** for actual hospital deployment

Remove patients without viable data

Extract and clean text to **prevent hindsight**



Input fields **mimic information available to nurses**

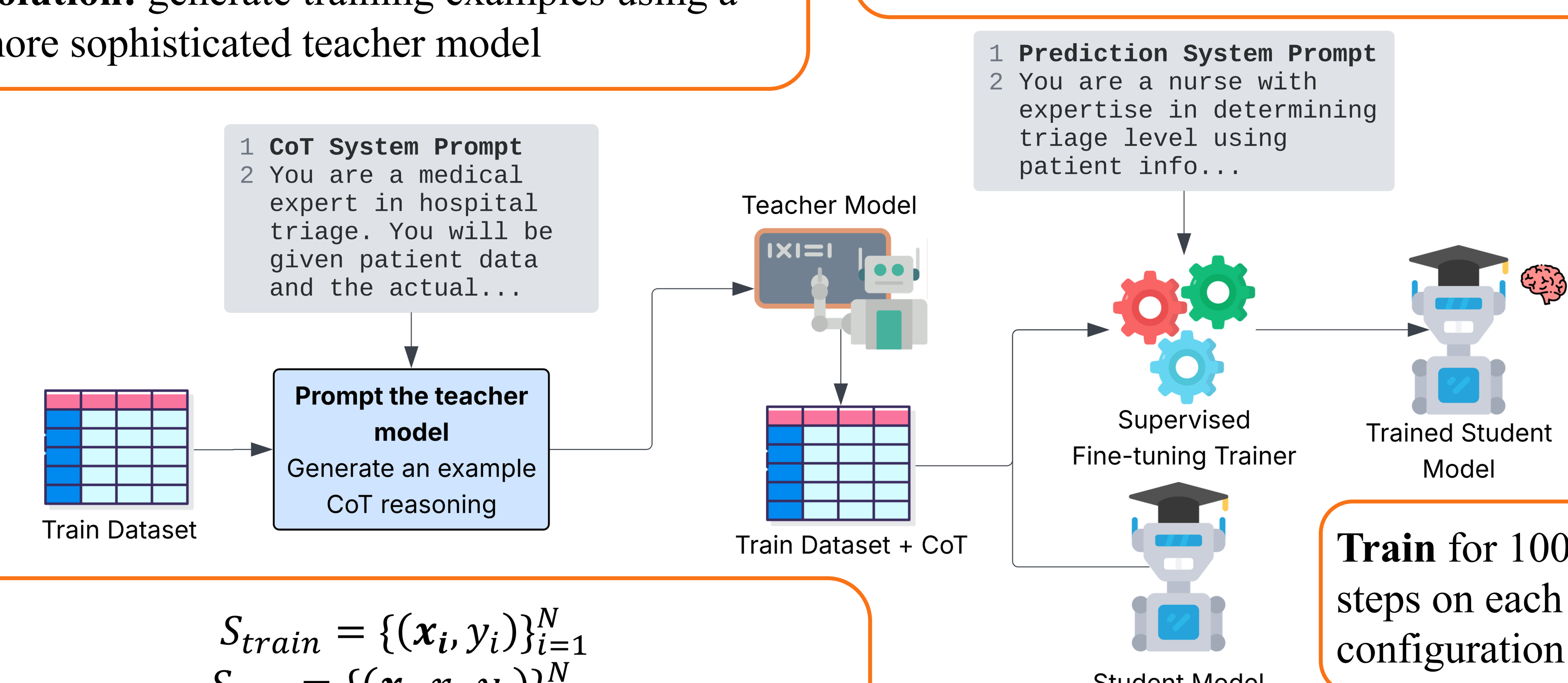
Use acuity field as triage ground truth

Train set: 12,000
Test set: 2,000

5. METHODS: MODEL DISTILLATION PIPELINE

Problem: dataset does not contain CoT examples, cannot finetune model
Solution: generate training examples using a more sophisticated teacher model

Prompts instruct model to think like a nurse: identify info, explain significance, compare options



Train for 1000 steps on each configuration

$$S_{train} = \{(x_i, y_i)\}_{i=1}^{N}$$
$$S_{CoT} = \{(x_i, r_i, y_i)\}_{i=1}^{N}$$

We provide the generated CoT example in addition to the input features and ground truth

CLINICAL VIABILITY

✓ Inference Time: <1 minute
✓ Accuracy: 62.68% (best model) vs 59% (humans)
✓ Deployment: can **run on most computers** with 16 GB RAM
✓ Privacy: data stays within the hospital

Use cases:
- Serve as secondary opinion
- Reduce **wait times**
- Flag cases for review

Paper Link

tinyurl.com/zhaotriage

* This poster includes additional figures and analyses not present in the submitted manuscript

6. EXPERIMENTS: ABLATION STUDY

🏆 **Beats GPT-5**
+3.83% accuracy
+0.08 κ

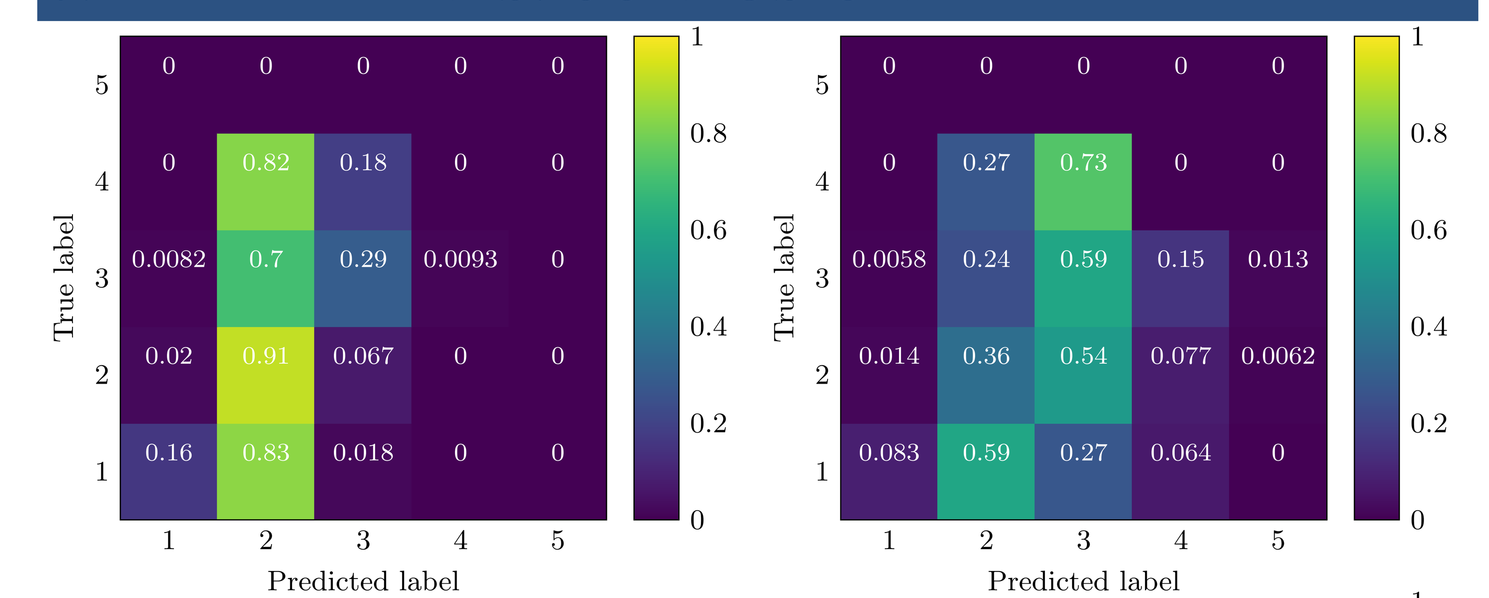
📈 **Significant Improvement**
+18.02% accuracy
+0.22 κ

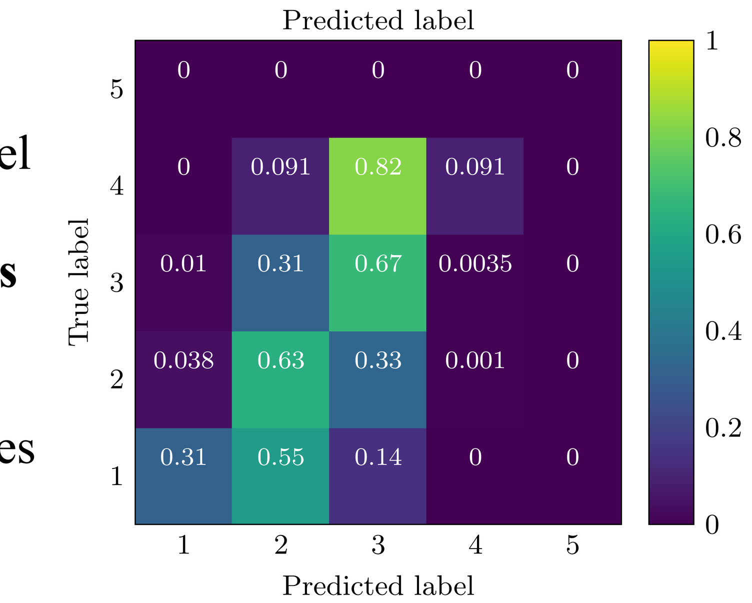
We evaluate seven **rank** (r), **alpha** (α), and **learning rate** (η) variations
Acc@1: raw accuracy of the models
$F1_{macro}$: evaluates performance on all triage classes (1-5) without regard to rarity
κ: quadratic weighted kappa penalizes larger ordinal errors

Model			Metrics		
r	α	η	Acc@1	$F1_{macro}$	κ
GPT-5			58.85	**84.70**	0.3270
gpt-oss-20b			44.66	45.89	0.1808
128	256	2e-4	60.42	60.11	0.3849
256	512	2e-4	62.51	62.11	0.4018
64	128	2e-4	60.71	60.17	0.3480
128	128	2e-4	60.81	60.37	0.3651
128	256	5e-5	57.29	56.81	0.3133
128	256	1e-4	61.07	60.67	0.3769
128	256	4e-4	**62.68**	62.36	**0.4056**

➢ Used κ to choose "best" model

7. EXPERIMENTS: CONFUSION MATRIX





Left to right: GPT-5, base gpt-oss-20b, best fine-tuned model
- GPT-5 prefers to over-triage
- Fine-tuned model exhibits **less spillover** than the base model
- Model tends to under-triage patients with more severe cases
- Errors mostly **within 1 level**

8. CONCLUSIONS AND FUTURE WORK

1. Built a **realistic** triage dataset using real patient records
2. Proposed a CoT distillation pipeline for creating light, open-weight medical models, **reducing reliance on proprietary systems**
3. Raised metrics by 15%+, beating GPT-5 and human performance

Limitations:
- Ground truths from real-world data may contain errors and are noisy
  - Independent expert verification, factor in agreement
- Does not account for unpredictable data input under time pressure
  - Study performance of model with real nurses