Privacy-First Triage Classification with Open-Weight LLMs

A Chain-of-Thought Distillation Approach

Zeyuan Zhao
Montgomery Blair High School
Silver Spring, Maryland

Yexiao He, Ang Li
University of Maryland
College Park, Maryland

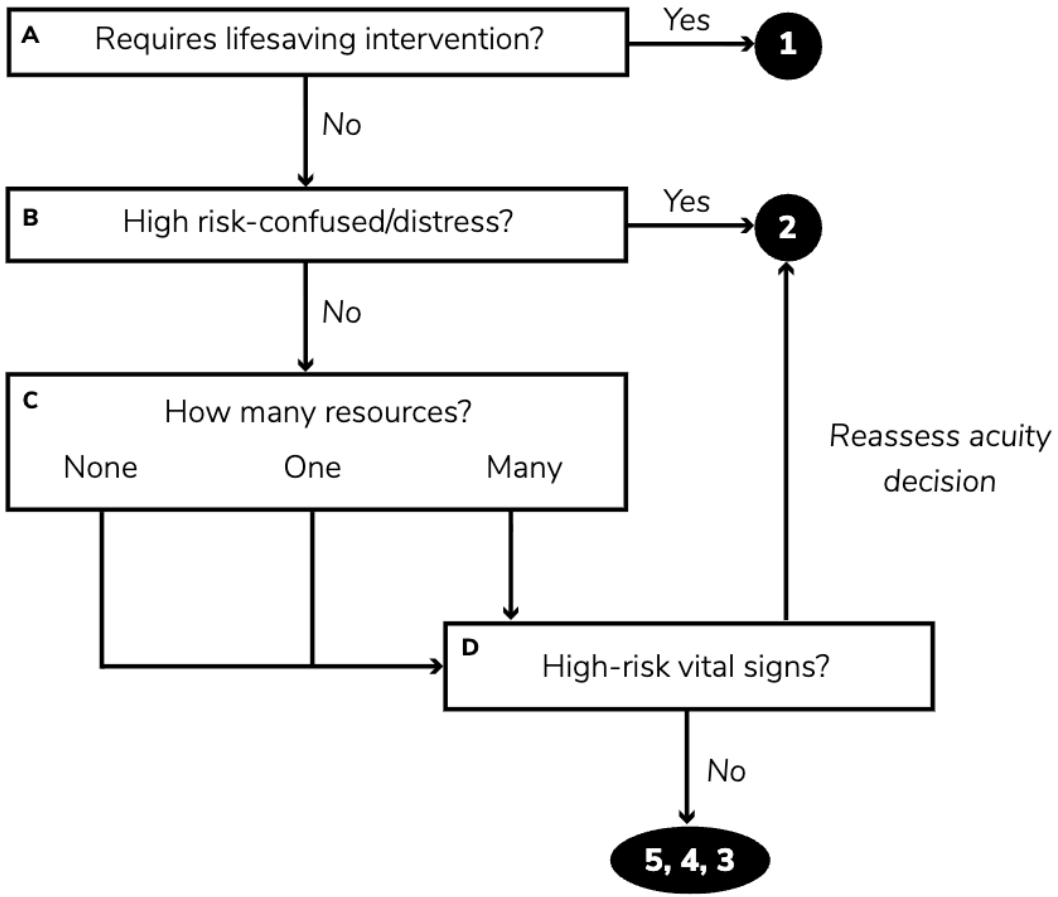
Background

Triage

Goal: prioritize severe cases, allocate resources efficiently

- Assess patients upon arrival to the Emergency Department (ED)
- Vitals, chief complaint, brief history, *etc*.
- Emergency Severity Index (ESI): 1 (most) – 5 (least severe)

ESI Flow



Source: Emergency Severity Index Handbook Fifth Edition

59%

Nurse Triage Accuracy*

*Source: Emergency Severity Index Handbook Fifth Edition

Large Language Models (LLMs)

- Current triage procedures have high error rates
- Skilled at reasoning and understanding free-form text
- May apply well in medical tasks

- Gaber *et al*. — Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis

Problem

- Proprietary systems, closed-weight
  - Privacy concerns
  - Transmit sensitive data over the internet
  - Expensive
  - Cannot host locally
  - Inaccessible in remote regions
- Open-weight models generally perform worse

Goal

Create a triage classifier model that is:

- Accurate
- Open-weight
- Small, deployable locally

… so that it…

- Handles real-world cases
- Protects patient privacy
- Reduces cost

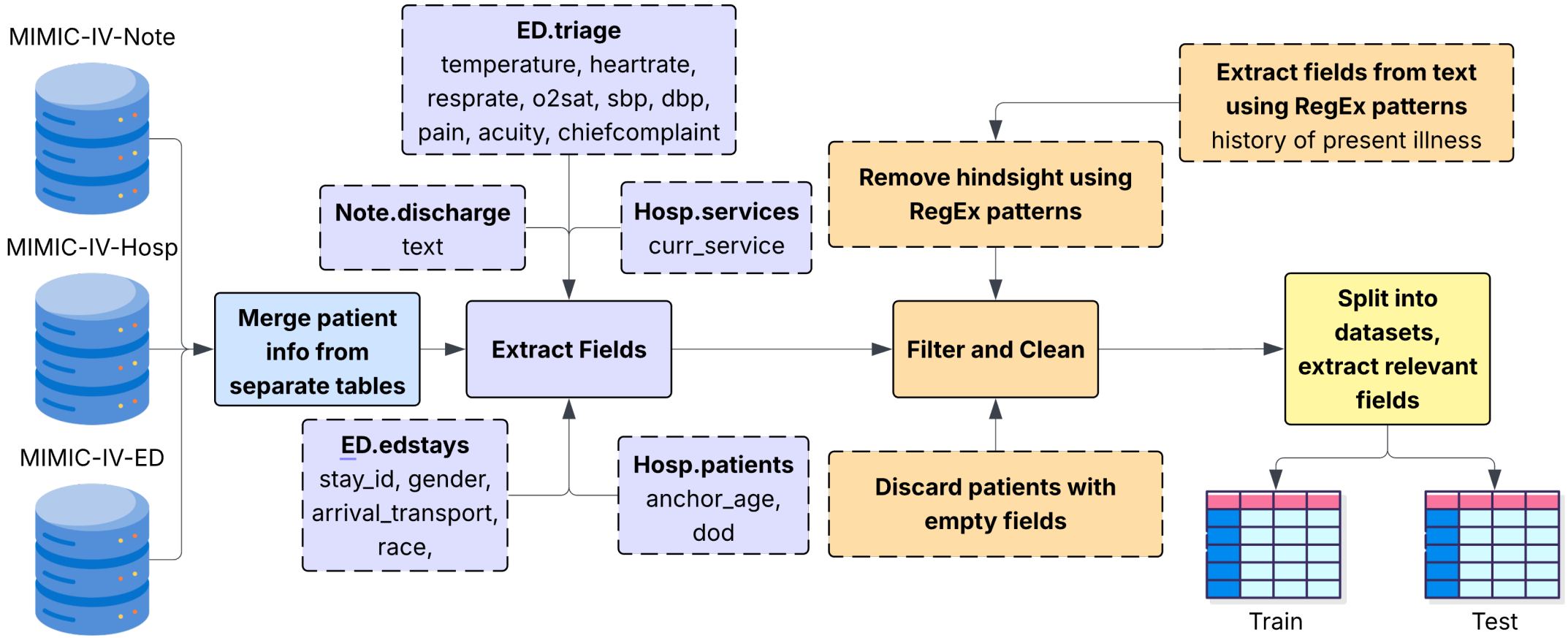
Methods

MIMIC-IV

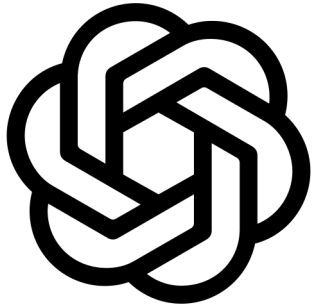
- Publicly available
- Hospital data
  - 200,000 ED patients
  - Triage data

Dataset Creation

Goal: simulate real triage scenarios



Model



Criteria

- Open-weight & Low Compute

OpenAI gpt-oss-20b

- HealthBench: 42.5%
- Chain-of-Thought (CoT)
- MXFP4 Quantized
- Mixture-of-Experts

Model Distillation

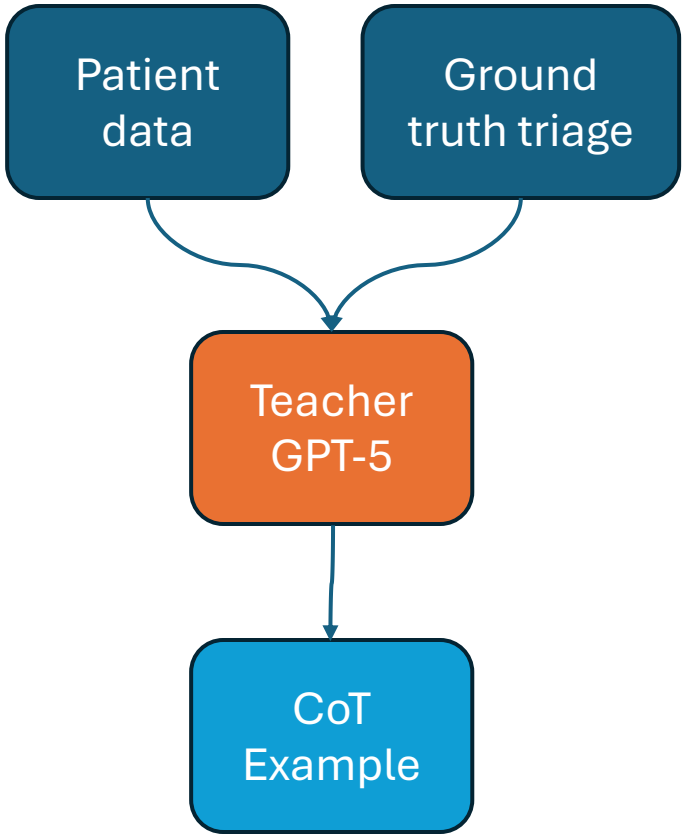
- Need CoT examples

Teacher Model Criteria

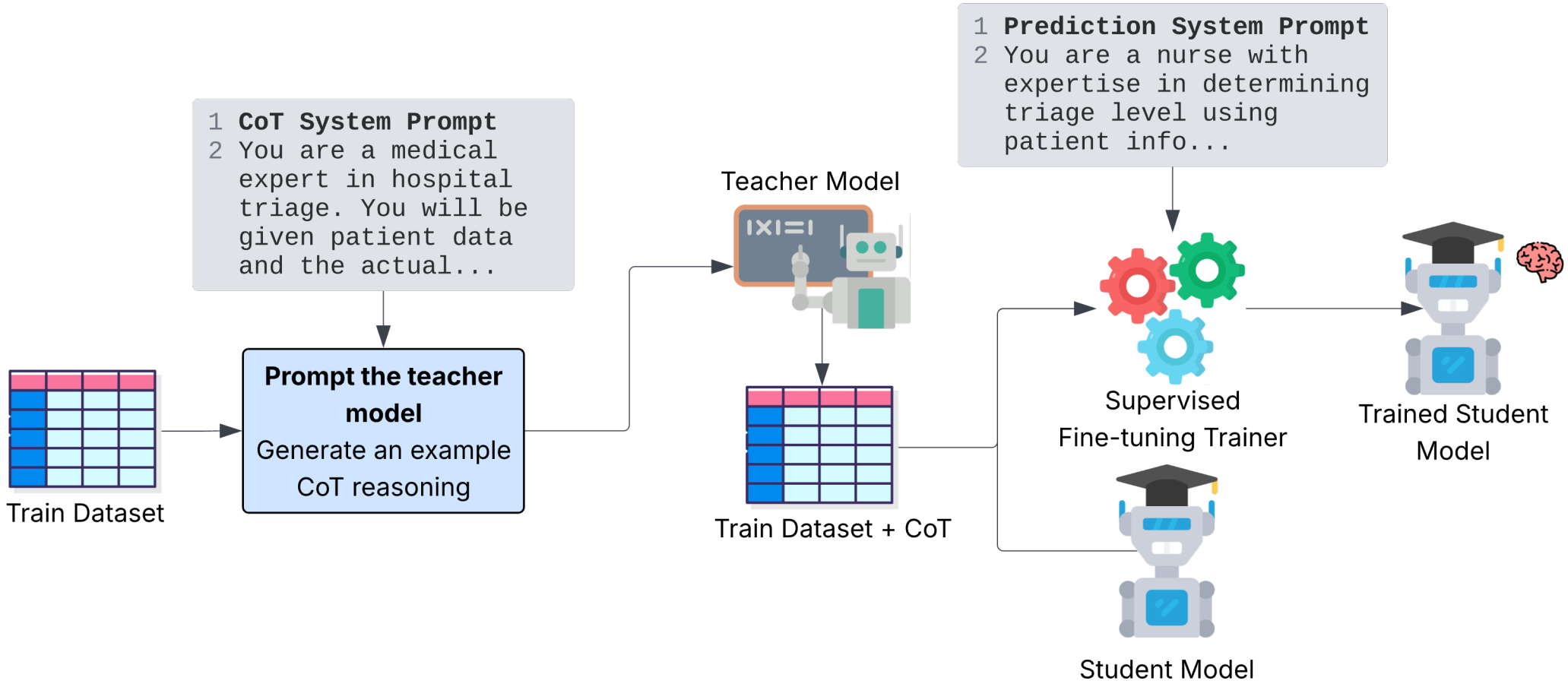
- High triage task performance

OpenAI GPT-5

- HealthBench: 67.2%



Model Distillation Pipeline

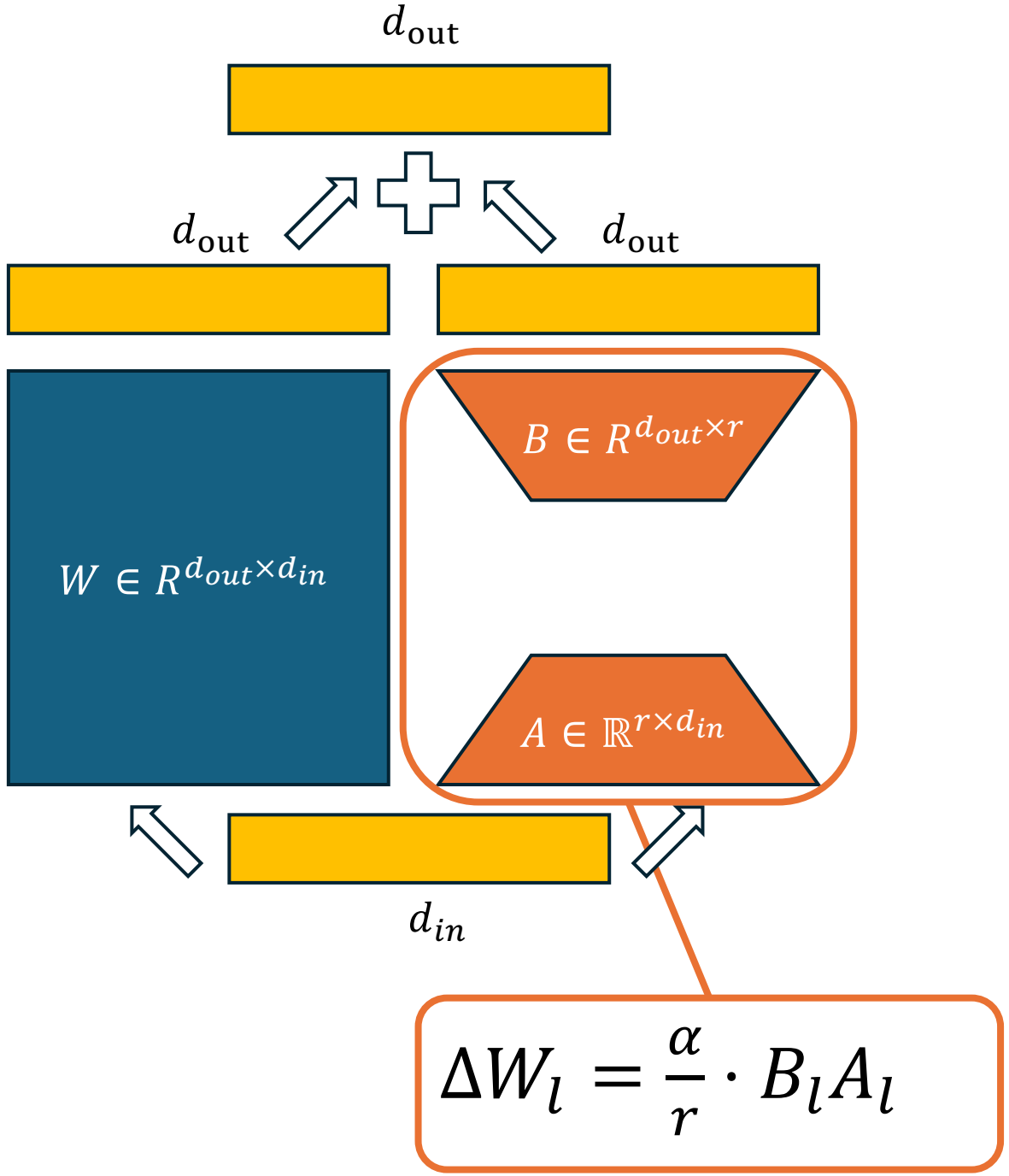


Low-Rank Adaptation

- Parameter-efficient fine-tuning
- Original weights frozen
- Adapters
  - MLP
  - Attention

α: scaling factor

r: rank of matrix decomposition



Summary

Chain-of-thought (CoT) distillation on an open-weight model

- Build dataset from real-world records
- Generate synthetic CoT
- Fine-tune base student model with low-rank adaptation

Experiments

Configuration

- 4x NVIDIA RTX 6000 Ada 48GB for training
- Quantized for inference
- Checkpoint evaluations: 250, 500, 1000

Ablation Study

9 models total:

- 6 finetune variations + 1 control
  - Rank
  - Alpha
  - Learning Rate
- Base gpt-oss-20b
- GPT-5

Hyperparameter	**Value**
Rank (r)	128
Alpha (α)	256 ($2r$)
Learning rate (η)	2e-4
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

Control training hyperparameters

Metrics

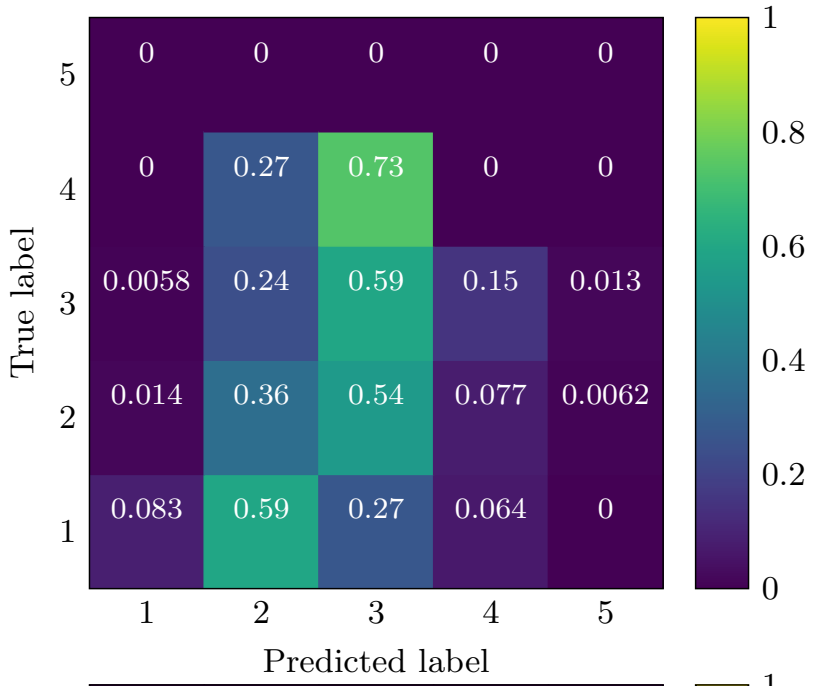
Model			Metrics		
r	α	η	Acc@1	$\mathrm{F1}_{\mathrm{macro}}$	κ
GPT-5			58.85	**84.70**	0.3270
gpt-oss-20b			44.66	45.89	0.1808
128	256	2e-4	60.42	60.11	0.3849
256	512	2e-4	<u>62.51</u>	62.11	<u>0.4018</u>
64	128	2e-4	60.71	60.17	0.3480
128	128	2e-4	60.81	60.37	0.3651
128	256	5e-5	57.29	56.81	0.3133
128	256	1e-4	61.07	60.67	0.3769
128	256	4e-4	**62.68**	<u>62.36</u>	**0.4056**

Use κ to choose “best”

Use κ to choose “best”

Confusion Matrix

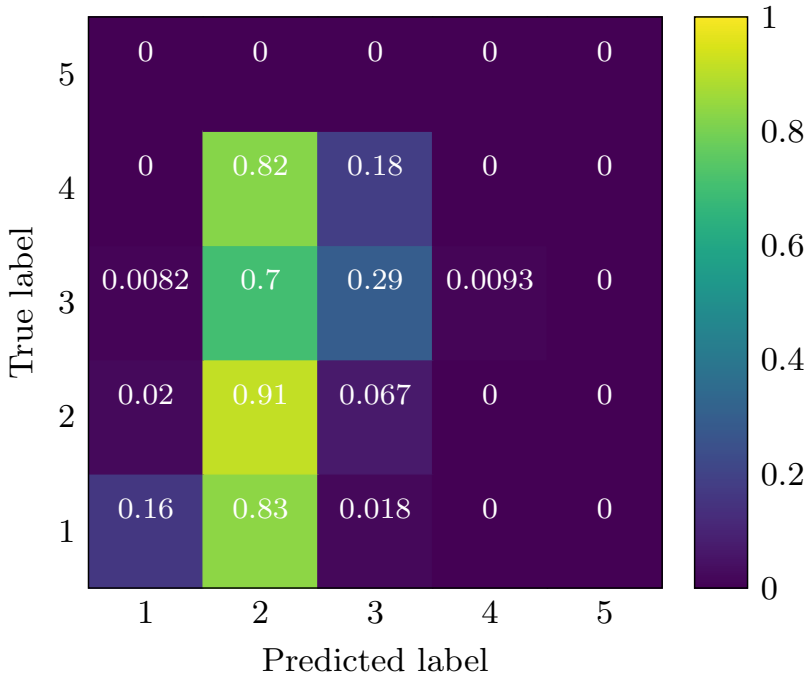
Spillover indicates inaccuracy

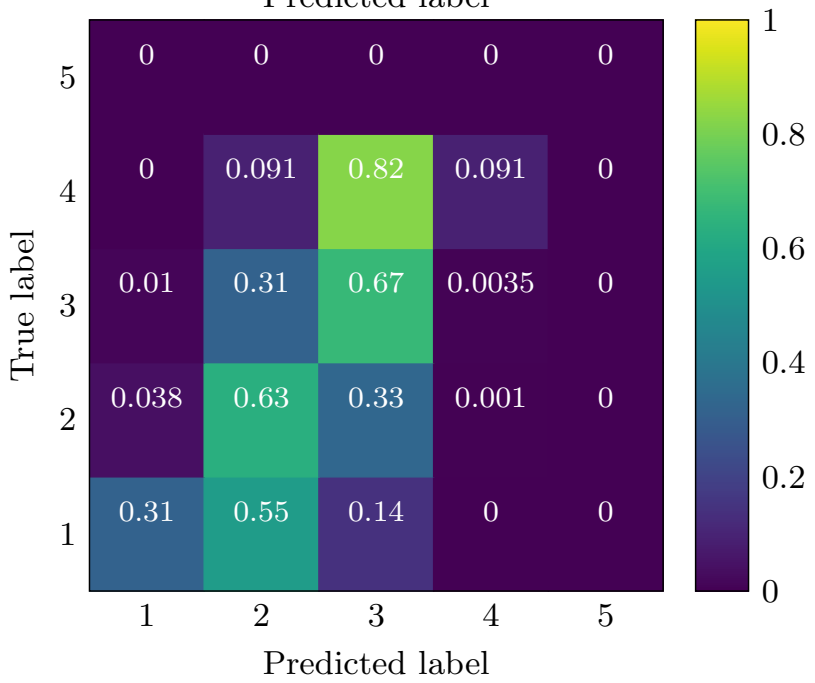


Control Base gpt-oss-20b

Lean left, severe over-triage

GPT-5





More accurate, clustering toward diagonal

Best Finetuned gpt-oss-20b

Implications

Clinical Viability

- Accuracy
  - Best finetuned model: 62.68%
  - Human accuracy: 59%
- Inference: <1 minute
- Memory: 16 GB
  - Local computers
- Privacy: no internet
  - No third-parties
  - Regulations

Impact

- Reduce costs
- Less reliance on proprietary systems
- Underserved areas
- Lower wait times

Use cases:

- Serve as secondary opinion
- Flag cases for review

Conclusion

1. Built a realistic triage dataset
   - Used real patient records
2. Proposed a CoT distillation pipeline
   - Created light, open-weight medical models
3. Beat GPT-5 and human performance
   - Raised metrics by 15%+ from base model

Future Work/Limitations

- MIMIC-IV contains real data
  - May contain errors and noise
  - Future: independent expert verification
- Unpredictable data input under time pressure
  - Nurses must input numbers and info
  - Incomplete data
  - Future: study performance in real simulations with nurses