ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance

Liwen Sun[1], Abhineet Agarwal[2], Aaron Kornblith[3], Bin Yu[2], Chenyan Xiong[1]
[1]Carnegie Mellon University, [2]University of California, Berkeley, [3]University of California, San Francisco

Agenda

- Motivation
- MIMIC-ED-Assist Benchmark
- ED-Copilot for Diagnostic Assistance
- Experiments
- Conclusion

Motivation: Emergency Department (ED) Crowding

A crucial factor affecting throughput is the **laboratory testing** process, where patients often face lengthy waits for tests to be ordered and completed, delaying diagnosis and treatment decisions.

MIMIC-ED-Assist Benchmark Objectives

- **Critical outcome:** if the patient is transferred to ICU or there is an inpatient mortality. Identifying patients with critical outcome allows clinicians to prioritize treatment and resources for them.

- **Lengthened ED Stay:** if the length of stay (LOS) exceeds 24 hours. Lengthened ED stay is typically correlated with the complexity of a patient's case.

MIMIC-ED-Assist Benchmark Curation

Data Preprocessing: Exclude patients that miss triage results and perform same tests multiple times.

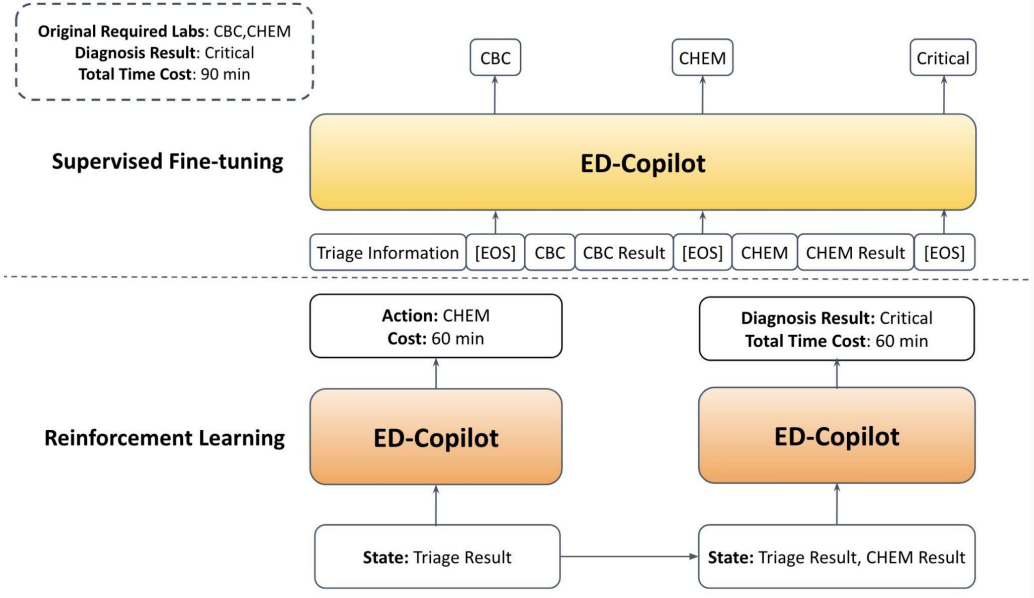
Triage Feature Selection: Include demographics, medical history, vital signs, and chief complaint.

Laboratory Test Selection: Categorize 67 available laboratory tests in ED into 12 distinct groups.

Variable/Label	Count
# of ED visits	32356
# of patients	25714
# of triage variables	9
# of laboratory variables	67
# of laboratory groups	12
Avg. # of laboratory groups per patient	4.7
# of Inpatient mortality	467 (1.44%)
# of ICU transfer in 12h	2894 (8.94%)
# of Critical outcome	3129 (9.67%)
# of ED LOS $>$ 24h	2232 (6.90%)

ED-Copilot for Diagnostic Assistance

We proposes a ED-Copilot system to offer (time) cost-effective diagnostic assistance by selecting informative tests and improving outcome for high-risk patients.



Preliminary

- As laboratory results and triage information are stored in a tabular format, we first linearize this information for PLM via textual template:

  test name : test value | test name : test value ……

- Apply PLM to obtain hidden representations for the text sequence and use two MLP on tokens [EOS] to predict the next laboratory group and outcome.

$$[x_0, r_0, \texttt{[EOS]}_0, \ldots, x_n, r_n, \texttt{[EOS]}_n, y] \xrightarrow{G_\theta} \boldsymbol{H},$$
$$p_\phi(x_i|\boldsymbol{h}_{<i}) = \text{MLP}_\phi(\boldsymbol{h}_{i-1}),$$
$$p_\psi(y|\boldsymbol{h}_{\leq n}) = \text{MLP}_\psi(\boldsymbol{h}_n).$$

Methodology (Stage 1)

Supervised Fine-tuning: To predict the next laboratory group and final outcome, we use a standard auto-regressive loss function. ED clinicians can use the fine-tuned PLM to suggest a sequence of laboratory groups and predict outcomes.

$$\mathcal{L}_{\text{lab}} = -\frac{1}{n}\sum_{i=1}^{n} \log p_\phi(x_i|\boldsymbol{h}_{<i}).$$

$$\mathcal{L}_y = -\log p_\psi(y|\boldsymbol{h}_{\leq n}).$$

Methodology (Stage 2)

Reinforcement Learning: We introduce the notion of time-cost effectiveness to the fine-tuned PLM by selecting laboratory groups that maximize predictive accuracy while minimizing time-cost.

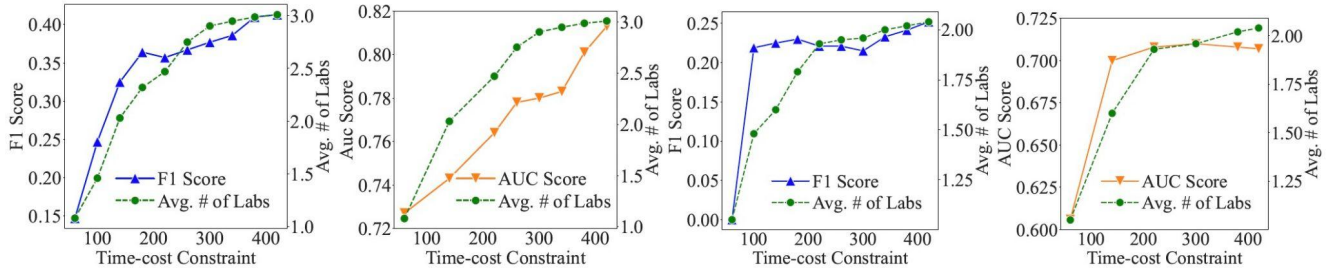
$$\text{Cost}(\pi_\eta) = \mathbb{E}_{\pi_\eta}[\sum_{t\geq 0}\sum_{j\in[K]} c(j)\cdot \mathbf{1}\{a_t = x_j\}],$$

$$\pi_\eta^*(\alpha,\beta) = \text{argmax}_{\pi_\eta}\{\text{TN}(\pi_\eta) + \alpha\text{TP}(\pi_\eta) + \beta\text{Cost}(\pi_\eta)\}.$$

Experiments: Overall Performance

Sensitivity and specificity are true positive and true negative rates. We report results averaged over three random seeds alongside standard deviations.

	Critical Outcome					Lengthened ED Stay				
Model	F1	AUC	Sensitivity	Specificity	Avg. Time-cost	F1	AUC	Sensitivity	Specificity	Avg. Time-cost
Random Forest	0.377 (0.015)	0.807 (0.011)	0.754 (0.012)	0.748 (0.005)	265 Min	0.206 (0.014)	0.698 (0.011)	0.693 (0.016)	0.616 (0.024)	265 Min
XGBoost	0.379 (0.019)	0.807 (0.009)	0.731 (0.017)	0.744 (0.006)	265 Min	0.212 (0.010)	0.679 (0.007)	0.619 (0.020)	0.661 (0.020)	265 Min
LightGBM	0.394 (0.016)	0.813 (0.008)	0.725 (0.012)	0.769 (0.004)	265 Min	0.217 (0.015)	0.705 (0.011)	0.706 (0.017)	0.605 (0.014)	265 Min
3-layer DNN	0.339 (0.032)	0.743 (0.021)	0.676 (0.024)	0.683 (0.011)	265 Min	0.194 (0.031)	0.637 (0.013)	0.649 (0.015)	0.593 (0.014)	265 Min
SM-DDPO	0.353 (0.031)	0.780 (0.020)	0.685 (0.023)	0.763 (0.022)	182 (32) Min	0.183 (0.028)	0.619 (0.012)	0.472 (0.012)	0.739 (0.011)	177 (60) Min
ED-Copilot	**0.413 (0.028)**	**0.820 (0.021)**	**0.750 (0.018)**	**0.779 (0.011)**	**125 (21)** Min	**0.232 (0.023)**	**0.707 (0.015)**	**0.725 (0.018)**	**0.606 (0.015)**	**154 (33)** Min



(a) Critical Outcome F1 (b) Critical Outcome AUC (c) Lengthened ED Stay F1 (d) Lengthened ED Stay AUC

Figure 2. Prediction accuracy and average number of laboratory groups of ED-Copilot with different maximum allowed time to perform laboratory tests. Each point reflects ED-Copilot's F1/AUC (y-axes) at different time upper-bounds.

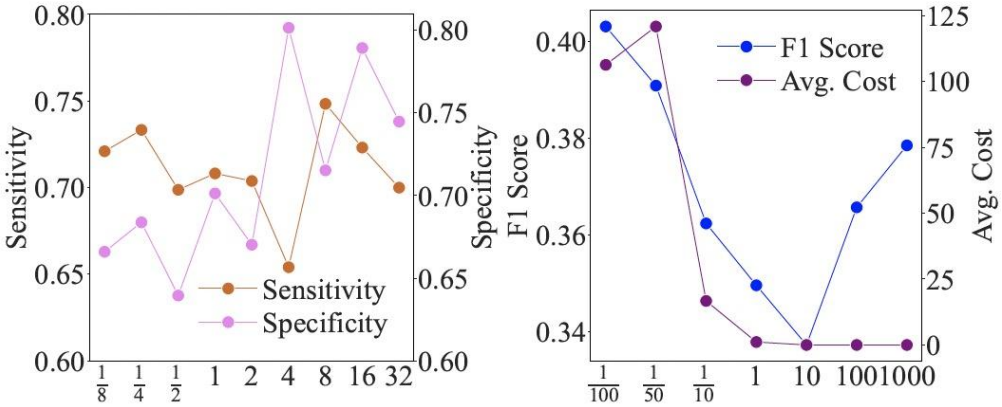
Experiments: Ablation Study

- Linearization Technique
- Feature Importance
- PLM Backbone

Group	Variations	Critical Outcome					Lengthened ED Stay				
		F1	AUC	Sensitivity	Specificity	Avg. Time-cost	F1	AUC	Sensitivity	Specificity	Avg. Time-cost
	ED-Copilot (345M)	**0.413**	**0.820**	**0.750**	**0.779**	**125 Min**	**0.252**	**0.707**	**0.725**	**0.606**	**154 Min**
Linearization	Raw Lab Test Name	0.397	0.777	0.768	0.677	134 Min	0.241	0.695	0.611	0.701	144 Min
Features	w/o. Triage	0.277	0.704	0.679	0.649	—	0.145	0.593	0.532	0.606	—
	w/o. CBC	0.385	0.803	0.692	0.777	—	0.224	0.686	0.696	0.596	—
	w/o. CHEM	0.420	0.827	0.788	0.746	—	0.234	0.702	0.656	0.606	—
Backbone	BioGPT (345M) w/o. RL	0.381	0.810	0.725	0.765	265 Min	0.236	0.718	0.710	0.620	265 Min
	Llama (7B LORA) w/o. RL	0.397	0.798	0.692	0.767	265 Min	0.232	0.701	0.705	0.610	265 Min
	Pythia (70M) w. RL	0.290	0.698	0.574	0.702	166 Min	0.178	0.596	0.555	0.619	126 Min
	GPT2-Medium (345M) w. RL	0.358	0.757	0.621	0.747	133 Min	0.166	0.539	0.498	0.584	96 Min

Experiments: Hyperparameter-control

- (α, β) control the trade-off between prediction accuracy and time-cost in training.
- Increasing α trades off sensitivity over specificity, while increasing β trades off F1-score over time-cost.



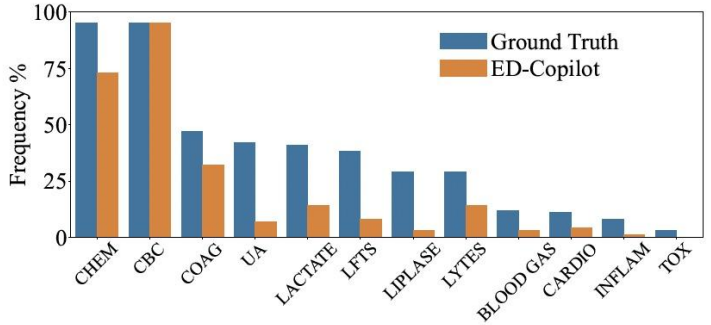
(a) Sensitivity-Specificity α (b) F1-Cost β

$$\pi_\eta^*(\alpha, \beta) = \text{argmax}_{\pi_\eta} \{\text{TN}(\pi_\eta) + \alpha\text{TP}(\pi_\eta) + \beta\text{Cost}(\pi_\eta)\}.$$

Experiments: Personalized Diagnostic Assistance

- We plot both the fraction of patients receiving each group of tests and the fraction of patients predicted by ED-Copilot. After the two most common groups (CHEM and CBC), more than half of the patients performed some other tests.

- On average each patient actually performed 4.7 groups and cost-effective ED-Copilot suggested 2.4 groups.



Experiments: Personalized Diagnostic Assistance

- We partition patients into three cohorts based on the rarity of laboratory groups they were administered to highlight the benefit of personalization from ED-Copilot.

- The middle and rare cohorts have higher severity (positive cases), ED-Copilot achieves significantly higher sensitivity than other methods.

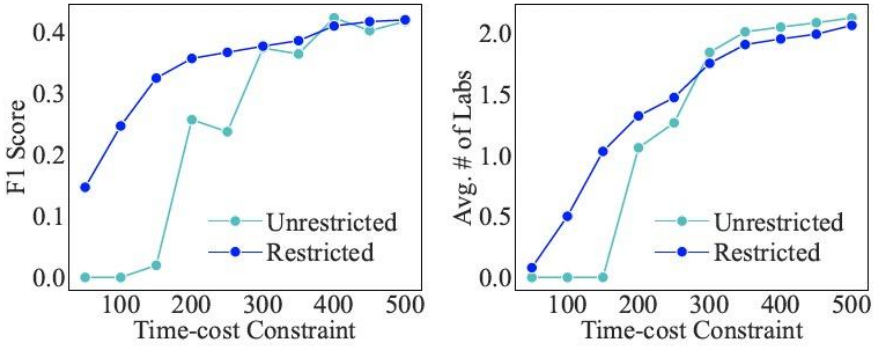
	W. Top 2 Lab Groups (302/2823,9.6%)			**W. Middle 6 Lab Groups** (299/2603, 10.3%)			**W. Last 4 Lab Groups** (141/817, 14.7%)		
Model	F1	Sensitivity	Specificity	F1	Sensitivity	Specificity	F1	Sensitivity	Specificity
Random Forest (Top 2 groups)	0.330	0.735	0.752	0.335	0.746	0.750	0.405	0.730	0.739
XGBoost (Top 2 groups)	0.361	0.788	0.680	0.374	0.732	0.728	0.433	0.738	0.718
LightGBM (Top 2 groups)	0.401	0.788	0.705	0.409	0.806	0.690	0.462	0.738	0.742
SM-DDPO	0.364	0.760	0.715	0.373	0.762	0.724	0.435	0.732	0.734
ED-Copilot	0.414	0.701	0.788	0.431	0.731	0.774	0.461	0.767	0.720

The total number of positive (critical)/negative cases and positive rate is shown in parentheses.

Experiments: Unrestricted Lab Group Suggestion

- Since MIMIC-ED-Assist is an offline retrospective benchmark, we restrict ED-Copilot during training to only select laboratory groups that patients have received.

- Without restriction to select observed laboratory tests (imputation by zero) for online evaluation, ED-Copilot achieved reasonable performance as the maximum allowed time and actual laboratory group increase.



(a) Accuracy (b) Avg. # of Labs

Conclusion

- We work with ED clinicians to develop MIMIC-ED-Assist, a publicly accessible benchmark designed to advance research in ED diagnostic assistance.

- We develop ED-Copilot, an cost-effective RL-trained language model that enhances ED workflow by suggesting informative laboratory groups, flagging high-risk patients and minimizing waiting time.

- Experiments demonstrate that ED-Copilot improves accuracy over state-of-the-art classifier while reducing ED wait-time by 50%.