CURA: Clinical Uncertainty Risk Alignment for Language Model–Based Risk Prediction

Sizhe Wang[1] **Ziqi Xu**[1] **Claire Najjuuko**[1]
Charles Alba[1] **Chenyang Lu**[1]
[1]Washington University in St. Louis
{sizhew, ziqixu, c.najjuuko, alba, lu}@wustl.edu

Abstract

Clinical language models (LMs) are increasingly applied to support clinical risk prediction from free-text notes, yet their uncertainty estimates often remain poorly calibrated and clinically unreliable. In this work, we propose **C**linical **U**ncertainty **R**isk **A**lignment (**CURA**), a framework that aligns clinical LM-based risk estimates and uncertainty with both individual error likelihoods and cohort-level ambiguities. CURA first fine-tunes domain-specific clinical LMs to obtain task-adapted patient embeddings, and then performs uncertainty fine-tuning of a multi-head classifier using a *bi-level* uncertainty objective. Specifically, an individual-level calibration term aligns predictive uncertainty with each patient's likelihood of error, while a cohort-aware regularizer pulls risk estimates toward event rates in their local neighborhoods in the embedding space and places extra weight on ambiguous cohorts near the decision boundary. We further show that this cohort-aware term can be interpreted as a cross-entropy loss with neighborhood-informed soft labels, providing a label-smoothing view of our method. Extensive experiments on MIMIC-IV clinical risk prediction tasks across various clinical LMs show that CURA consistently improves calibration metrics without substantially compromising discrimination. Further analysis illustrates that CURA reduces overconfident false reassurance and yields more trustworthy uncertainty estimates for downstream clinical decision support. Our code is available at: https://github.com/sizhe04/CURA.

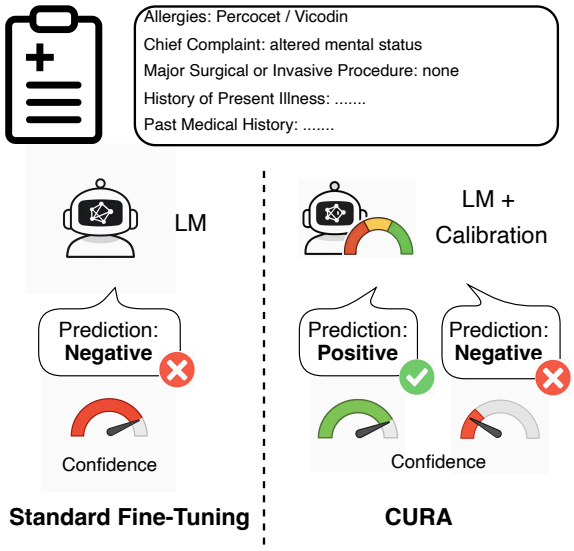

Figure 1: Comparison between standard fine-tuning and our proposed CURA framework in clinical risk prediction. A vanilla fine-tuned clinical LM (left) may produce overconfident wrong predictions, while CURA (right) encourages high confidence only for correct decisions and assigns higher uncertainty to potentially erroneous predictions.

1 Introduction

Clinical language models (LMs) (Alsentzer et al., 2019; Huang et al., 2019; Luo et al., 2022b) have demonstrated remarkable capabilities in processing unstructured electronic health records (EHRs) to support clinical decision-making (Jiang et al., 2023; Yang et al., 2022). By fine-tuning on free-text clinical notes, these models can effectively extract patient representations to predict outcomes such as mortality risk and length of stay (Harutyunyan et al., 2019; Johnson et al., 2023b). However, a key barrier to deploying such models in safety-critical settings is the reliability of their uncertainty estimates. While recent advancements have pushed the boundaries of discriminative performance (Nori et al., 2023; Singhal et al., 2025), overconfident yet incorrect predictions still pose a direct danger to patient safety.

In clinical scenarios, reliable uncertainty quantification is crucial alongside prediction accuracy. A well-calibrated model should act as a transparent partner, signaling high uncertainty for ambiguous or out-of-distribution cases to trigger expert review, while automating low-risk, high-confidence predictions. However, although fine-tuning typ-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34000–34018

ically improves predictive performance, it often exacerbates the model's tendency toward overconfidence (Guo et al., 2017), leading to cases where the model is highly confident yet wrong on high-risk patients as shown in Figure 1 (left). To mitigate this issue, various calibration techniques have been proposed. General uncertainty estimation methods, such as Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) estimate uncertainty by aggregating predictions from multiple stochastic or independently trained models. In parallel, recent works have explored calibration strategies tailored for generative LLMs, often leveraging verbalized probabilities or prompting techniques (Wang et al., 2024a; Leng et al., 2024; Huang et al., 2025a,b).

Nevertheless, directly applying these approaches in clinical settings remains challenging. First, general ensemble-based methods typically aggregate multiple predictions without leveraging the semantic structure in the representation space, adjust confidence scores only at the level of isolated samples, and may remain miscalibrated on borderline cohorts. Second, many LLM-specific calibration methods rely on expert-written rationales, high-quality chain-of-thought (CoT) annotations, or strong teacher models to refine verbalized uncertainty (Yang et al., 2024). However, in routine clinical workflows, downstream tasks are typically framed as binary risk stratification with only outcome labels available, and ground-truth explanations are rarely recorded at scale. Moreover, for complex patient profiles, even state-of-the-art LLMs may produce clinical reasoning that is difficult to trust, especially if such explanations are used as supervision targets (Nori et al., 2023). These constraints highlight the need for uncertainty calibration methods that operate directly on learned representations and binary labels, without relying on explicit textual rationales.

To address these limitations, we propose **CURA** (Clinical Uncertainty Risk Alignment), a bi-level (individual- and cohort-level) uncertainty calibration framework built on top of fine-tuned clinical LMs for risk prediction. As shown in Figure 1 (right), CURA aims to reserve high confidence for correct predictions while assigning higher uncertainty to potentially erroneous decisions. Specifically, our approach first fine-tunes a domain-specific clinical LM on free-text notes using standard cross-entropy to obtain task-adapted patient embeddings. On top of these frozen representations, we then train a multi-head classifier with a bi-level uncertainty objective that simultaneously aligns uncertainty at the individual and cohort levels. Additionally, this objective can be implemented as a lightweight plug-in loss on top of standard clinical LM fine-tuning, without relying on external explanations or multiple inference passes. Extensive experiments on MIMIC-IV clinical risk prediction tasks show that CURA consistently improves calibration metrics while maintaining—or slightly improving—discriminative performance, and further analysis demonstrates that CURA reshapes the uncertainty distribution, reduces false reassurance for high-risk patients, and provides greater practical utility for real-world clinical deployment.

To summarize, our contribution in this work is threefold:

- We propose CURA, a bi-level uncertainty calibration framework that calibrates clinical LMs by jointly aligning individual predictive uncertainty with cohort-level risk via neighbors in the embedding space.
- We formulate CURA as a lightweight plug-in objective that requires no architectural changes and admits an interpretation as cross-entropy with neighborhood-informed soft labels.
- Extensive experiments and analysis show that CURA not only improves calibration metrics but also yields trustworthy uncertainty estimates for safe clinical triage.

2 Related Work

Clinical LMs for Risk Prediction Recent work has shown that domain-specific clinical LMs, from BERT-style encoders (Alsentzer et al., 2019; Huang et al., 2019; Lee et al., 2020) to generative LLMs such as BioGPT (Luo et al., 2022b), can substantially improve risk prediction from free-text notes and other electronic health record (EHR) data (Singhal et al., 2022; Yang et al., 2022; Jiang et al., 2023; Alba et al., 2025).

However, the vast majority of these studies are primarily optimized and evaluated on discriminative metrics, with little attention to the calibration of the resulting risk estimates (Harutyunyan et al., 2019; Zhu et al., 2024; Chen et al., 2025a). Our work distinguishes itself by proposing a framework

that is designed to substantially improve the calibration of clinical LM–based risk prediction while maintaining, and in many cases slightly improving, discriminative performance.

Uncertainty Calibration for LLMs Uncertainty quantification is fundamental for reliable AI deployment in safety-critical domains such as healthcare (Kendall and Gal, 2017; Abdar et al., 2021; Gawlikowski et al., 2023; Xu et al., 2025). Classic approaches approximate Bayesian inference through techniques such as MC dropout and deep ensembles, while post-hoc calibration methods adjust confidence scores using temperature scaling or related transformations (Guo et al., 2017; Kull et al., 2019). In the era of LLMs, research has shifted toward eliciting uncertainty via verbalized confidence, self-evaluation, or semantic consistency (Lin et al., 2022; Tian et al., 2023; Kadavath et al., 2022; Melo et al., 2024; Kapoor et al., 2024). Beyond these generation-centric approaches, LLM-based reward models adopt multi-head ensembles on a shared backbone to obtain diverse uncertainty estimates at low computational cost (Liang et al., 2022; Duan et al., 2025). While promising for open-ended question answering and reasoning benchmarks, these generation-centric methods' calibration behavior in classification tasks remains largely unexplored (Liu et al., 2025) and is susceptible to overconfidence on distribution-shifted inputs (Zhou et al., 2023; Zulfiqar et al., 2025). Crucially, most prior uncertainty quantification techniques operate on individual samples in isolation, neglecting the structural information carried by similar examples. Our work bridges this gap by explicitly aligning predictive uncertainty with both individual error rates and cohort neighborhood risks in the embedding space, in a white-box setting where the clinical LM and prediction heads are fully accessible, rather than via prompt-based black-box access.

3 Methodology

In this section, we elaborate on the Clinical Uncertainty Risk Alignment (CURA) framework, depicted in Figure 2, which consists of two primary components: Section 3.1 describes fine-tuning domain-specific clinical LMs, and Section 3.2 presents the uncertainty fine-tuning.

3.1 Clinical LM Fine-Tuning

We address clinical risk prediction tasks using a dataset $\mathcal{D} = \{(x_i, y_i)\}$, where each input x_i consists of a de-identified clinical note, and the target $y_i \in \{0, 1\}$ serves as a binary indicator for the downstream task. We initialize our framework with a domain-specific clinical LM π_0 and attach a linear classifier on top.

Let π_θ denote the LM parameterized by θ, and let $p_\theta(x)$ be the predicted positive-class probability produced by the linear classifier on top of π_θ. We jointly optimize the parameters of the language model π_θ and the linear classifier by minimizing a class-weighted binary cross-entropy (Xie and Manski, 1989) loss $\mathrm{CE}(p_\theta(x), y)$ over $\mathcal{D}$. After fine-tuning, we freeze the updated LM π_θ and use it as a feature extractor to obtain fixed-dimensional embeddings $e_i = \pi_\theta(x_i)$ for all samples. These frozen embeddings serve as the input for the subsequent uncertainty fine-tuning step.

3.2 Uncertainty Fine-Tuning

Our proposed objective aligns three complementary aspects: (1) L_{base} ensures predictive accuracy against ground-truth labels (Section 3.2.1); (2) L_{ind} aligns the model's internal uncertainty with its own estimated error proxy (Section 3.2.2); and (3) L_{coh} regularizes predictions via local consistency with cohort risks (Section 3.2.3). Throughout, we use the term *bi-level* to refer to these individual- and cohort-level components of uncertainty alignment, rather than a nested bilevel optimization problem.

3.2.1 Classifier and Base Risk Loss

We construct a multi-head framework with an ensemble of M independent, randomly initialized lightweight MLP heads on top of the frozen embeddings $\{e_i\}$. At inference, we average predictions across all M heads. This architecture encourages diverse risk prediction and stabilizes calibration while sharing a single backbone for minimal inference cost (Duan et al., 2025). Given a patient note x_i with embedding $e_i = \pi_\theta(x_i)$, the ensemble average prediction is $\bar{p}(x_i) = \frac{1}{M}\sum_{m=1}^{M} p_m(e_i)$, and the base objective optimizes the expected cross-entropy between this mean prediction and the target y:

$$L_{base} = \mathbb{E}_{(x_i,y_i)}\big[\mathrm{CE}(\bar{p}(x_i), y_i)\big], \qquad (1)$$

where $\mathrm{CE}(\cdot,\cdot)$ is the same class-weighted loss used in the clinical LM fine-tuning step. This base loss provides the discriminative backbone for our uncertainty fine-tuning.

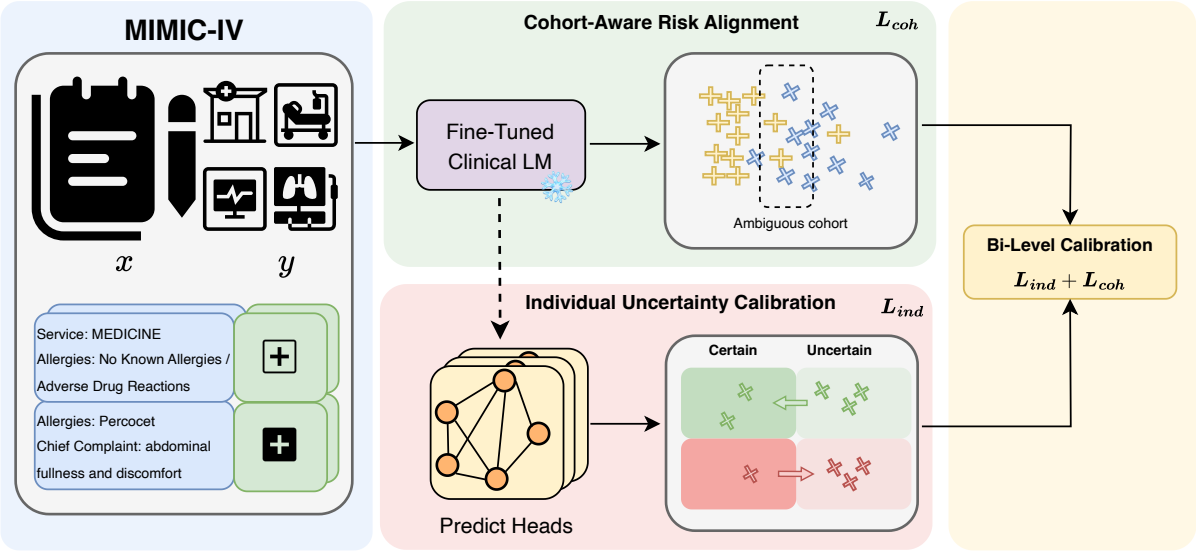

Figure 2: Overview of the CURA pipeline. The clinical LM is fine-tuned on MIMIC-IV notes to produce task-adapted embeddings for a multi-head classifier, trained with a base loss L_{base} plus a bi-level uncertainty calibration: the individual term L_{ind} calibrates patient-level uncertainty, and the cohort-aware term L_{coh} aligns predictions with neighborhood risks in the embedding space, emphasizing ambiguous cohorts.

3.2.2 Individual Uncertainty Calibration

While L_{base} ensures accurate risk estimation, it does not explicitly constrain how the model's confidence relates to its errors, so its predicted probability can still be poorly calibrated. Motivated by prior work on uncertainty-aware fine-tuning for LLMs (Krishnan et al., 2024), which is grounded in decision theory (Murphy, 2012) and couples token-level predictive entropy with correctness in inference, we introduce an analogous individual-level calibration term for binary risk prediction. For a patient note x, we define the correctness probability $a(x) = y\,\bar{p}(x) + (1-y)(1-\bar{p}(x))$, representing the confidence assigned to the ground-truth label. Based on information-theoretic views of predictive uncertainty, where entropy is used as a measure of distributional uncertainty (Shannon, 1948; Houlsby et al., 2011), we quantify the model's uncertainty via the entropy of the ensemble-averaged prediction:

$$H(x) = -\Big[\bar{p}(x)\log\bar{p}(x) + \big(1-\bar{p}(x)\big)\log\big(1-\bar{p}(x)\big)\Big]. \tag{2}$$

We normalize this entropy by its maximum possible value and obtain a scalar uncertainty score $u(x) = H(x)/H_{\max} \in [0,1]$, and align it with the error proxy $(1-a(x))$ by minimizing the individual-level calibration loss:

$$L_{ind} = \lambda_{ind}\,\mathbb{E}_{(x,y)}\Big[-(1-a(x))\log u(x) - a(x)\log\big(1-u(x)\big)\Big], \tag{3}$$

where λ_{ind} is a scaling hyperparameter. Minimizing L_{ind} aligns the model's uncertainty with its error rate at the individual patient level: confident and correct predictions are rewarded with small loss, whereas confident mistakes incur a large penalty. Notably, the L_{base} in Equation 1 acts as an anchor for discrimination, preventing the degenerate solution where the model outputs uniform probabilities to minimize L_{ind} trivially. Furthermore, minimizing L_{ind} effectively reshapes the probability distribution, penalizing high confidence specifically on samples where the model is prone to error. In all subsequent analyses, we use the same normalized entropy $u(x)$ computed from each method's final predicted probability as the scalar uncertainty score, so that all uncertainty-based comparisons are made under a common definition.

3.2.3 Cohort-Aware Risk Alignment

To extend risk alignment from the individual to the group level, we propose a *cohort-aware* term that ensures clinically similar patients receive consistent risk estimates. For each patient note x_i with embedding e_i, we retrieve its K nearest neighbors, denoted as $\mathcal{N}_K(e_i)$, among all labeled training sam-

ples and compute the event rate in this local cohort:

$$q(x_i) = \frac{1}{K} \sum_{j \in \mathcal{N}_K(e_i)} y_j. \tag{4}$$

We interpret $q(x_i)$ as the *neighborhood risk*, representing the inherent risk for patients with similar clinical presentations. To quantify cohort ambiguity, we compute the normalized entropy $\hat{H}(q(x_i))$, which increases when the cohort exhibits a mixed outcome distribution—a characteristic of borderline cases. We leverage this cohort information to regularize the model via a cohort-aware loss:

$$L_{coh} = \mathbb{E}_{x_i}\left[w(x_i)\,\mathrm{CE}\big(\bar{p}(x_i), q(x_i)\big)\right], \tag{5}$$

where the adaptive weight $w(x_i) = \lambda_{coh}\,\hat{H}\big(q(x_i)\big)$ is controlled by the hyperparameter λ_{coh}. This objective aligns $\bar{p}(x_i)$ with the neighborhood risk $q(x_i)$ and assigns larger training weights to cohorts with high neighborhood entropy.

Overall, our uncertainty fine-tuning objective is:

$$L_{total} = L_{base} + L_{ind} + L_{coh}. \tag{6}$$

Here L_{base} provides standard risk supervision for the multi-head classifier, while L_{ind} and L_{coh} together form the bi-level uncertainty regularizer illustrated in Figure 2. In Appendix F, we further show that minimizing $L_{base} + L_{coh}$ is equivalent to optimizing a single cross-entropy loss with a cohort-aware soft label, which can be viewed as a data-dependent label-smoothing scheme whose target interpolates between the ground-truth label and the neighborhood risk. This soft-label view highlights that cohorts with high neighborhood entropy pull the target toward their local event rate and thus receive stronger regularization, while in practice L_{total} can be used as a drop-in training objective on top of existing clinical LMs without modifying their inference pipeline. Because L_{ind} and L_{coh} are computed from model probabilities and pre-computed neighborhoods, CURA does not introduce additional backbone parameters or extra forward passes beyond the shared single-model pipeline. An empirical runtime comparison is provided in Appendix I.

4 Experiment

4.1 MIMIC-IV Dataset

We utilize de-identified clinical notes from the MIMIC-IV database (Johnson et al., 2023b) to predict five distinct risk stratification tasks. For each task, the input x is a single free-text note and the output $y \in \{0, 1\}$ indicates a pre-defined adverse outcome within a fixed horizon. These tasks align with benchmarks in prior clinical LM research (Xue et al., 2022; Luo et al., 2022a; Alba et al., 2025) and cover diverse clinical scenarios, ranging from short-term to medium-term prognosis. Specifically, we evaluate on: (1) 7-day mortality, (2) 30-day mortality, (3) early discharge within 12 hours, (4) in-hospital mortality, and (5) ICU stay longer than one day. Further details regarding the task definitions are provided in Appendix A.

4.2 Experiment Settings

Clinical Language Models We utilize three domain-specific language model backbones: BioGPT (Luo et al., 2022b), BioClinicalBERT (Alsentzer et al., 2019), and ClinicalBERT (Huang et al., 2019). For brevity, we present results for BioGPT in the main experiment, and defer the BioClinicalBERT and ClinicalBERT results to Appendix B.

Baselines Following prior work on uncertainty quantification for clinical outcome prediction (Chen et al., 2025b), we include two widely used architecture-agnostic white-box baselines: Deep Ensembles and MC dropout. In addition, we also consider an internal baseline that uses the same multi-head architecture as CURA but optimizes only the class-weighted cross-entropy in the uncertainty fine-tuning step. All baselines employ the same fine-tuned backbone models as our approach, so that any performance differences stem only from the choice of uncertainty mechanism. We also report a runtime comparison under identical hardware and early-stopping settings in Appendix I to evaluate the practical overhead of each uncertainty method. Further implementation details are provided in Appendix C.

Evaluation Metrics We evaluate model performance using two categories of metrics: (1) Discrimination: Area Under the ROC Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC); and (2) Calibration & Uncertainty: Brier score, Negative Log-Likelihood (NLL), and Area Under the Risk-Coverage curve (AURC). All reported results represent the mean and standard deviation across five cross-validation folds.

Task	Method	AUROC ↑	AUPRC ↑	Brier ↓	NLL ↓	AURC ↓
7-day Mortality	Baseline	0.852 (0.025)	0.127 (0.015)	0.032 (0.005)	0.120 (0.015)	0.008 (0.001)
	Deep Ensemble	0.856 (0.022)	0.125 (0.010)	0.029 (0.004)	0.110 (0.009)	0.007 (0.000)
	MC Dropout	0.862 (0.019)	0.130 (0.008)	0.034 (0.006)	0.127 (0.017)	0.009 (0.002)
	CURA (Ours)	**0.892** (0.010)	**0.160** (0.010)	**0.015** (0.003)	**0.075** (0.010)	**0.002** (0.000)
30-day Mortality	Baseline	0.881 (0.009)	0.270 (0.009)	0.064 (0.003)	0.231 (0.004)	0.024 (0.003)
	Deep Ensemble	0.869 (0.008)	0.256 (0.011)	0.060 (0.001)	0.215 (0.004)	0.020 (0.002)
	MC Dropout	0.880 (0.014)	0.268 (0.014)	0.063 (0.002)	0.230 (0.007)	0.024 (0.003)
	CURA (Ours)	**0.890** (0.004)	**0.280** (0.012)	**0.038** (0.001)	**0.146** (0.003)	**0.009** (0.001)
In-hospital Mortality	Baseline	0.621 (0.011)	0.027 (0.002)	0.044 (0.004)	0.175 (0.013)	0.015 (0.001)
	Deep Ensemble	0.632 (0.010)	0.028 (0.002)	0.044 (0.004)	0.171 (0.010)	0.015 (0.001)
	MC Dropout	0.629 (0.013)	0.028 (0.002)	0.046 (0.005)	0.174 (0.013)	0.015 (0.001)
	CURA (Ours)	**0.641** (0.009)	**0.029** (0.002)	**0.029** (0.004)	**0.124** (0.010)	**0.011** (0.000)
ICU Stay - 1	Baseline	0.579 (0.026)	**0.035** (0.003)	0.105 (0.042)	0.350 (0.108)	0.064 (0.053)
	Deep Ensemble	0.579 (0.024)	**0.035** (0.002)	0.104 (0.043)	0.345 (0.110)	0.062 (0.052)
	MC Dropout	0.580 (0.027)	**0.035** (0.003)	0.106 (0.043)	0.350 (0.113)	0.066 (0.056)
	CURA (Ours)	**0.584** (0.029)	**0.035** (0.003)	**0.085** (0.044)	**0.288** (0.104)	**0.040** (0.030)
Early Discharge - 12	Baseline	0.587 (0.011)	0.027 (0.012)	0.018 (0.002)	0.085 (0.006)	0.007 (0.001)
	Deep Ensemble	0.586 (0.012)	0.028 (0.012)	0.017 (0.001)	0.085 (0.005)	0.007 (0.000)
	MC Dropout	0.588 (0.013)	0.028 (0.012)	0.018 (0.002)	0.083 (0.008)	0.007 (0.000)
	CURA (Ours)	**0.594** (0.017)	**0.031** (0.014)	**0.010** (0.001)	**0.056** (0.003)	**0.006** (0.000)

Table 1: Performance of **BioGPT** on five clinical risk prediction tasks. Results are reported as mean (standard deviation) across five-fold cross-validation. "ICU Stay - 1" refers to ICU stays longer than one day, and "Early Discharge - 12" refers to early discharge within 12 hours.

4.3 Experimental Results

In this section, we compare our overall CURA method with the baseline and the two aforementioned uncertainty quantification strategies. As shown in Table 1, CURA achieves modest but stable improvements in discrimination performance (AUROC and AUPRC) relative to the baseline, indicating that our method does not sacrifice predictive accuracy. While Deep Ensembles and MC Dropout remain competitive in discriminative tasks, their contribution to calibration appears less pronounced in this setting. These methods yield only marginal reductions in Brier score, NLL, and AURC, and in some instances even slightly degrade these metrics. Comparatively, CURA exhibits stronger calibration capabilities, yielding consistent reductions across all three calibration metrics on all five tasks. We attribute this effectiveness to the synergistic effect of our individual and cohort-level alignment, which allows the model to better distinguish between reliable and ambiguous predictions. Overall, these results suggest that our bi-level uncertainty fine-tuning objective is particularly effective at reshaping the probability outputs of clinical LMs into reliable risk estimates, while retaining competitive discriminative performance. Similar robust trends are observed for BioClinicalBERT and ClinicalBERT, as detailed in Appendix B.

Method	AUROC	AUPRC	Brier	NLL	AURC
7-day Mortality					
Baseline	0.852	0.127	0.032	0.120	0.008
w/ L_{ind}	**0.892**	**0.160**	0.019	0.077	0.003
w/ L_{coh}	0.851	0.124	0.023	0.118	0.007
CURA	**0.892**	**0.160**	**0.015**	**0.075**	**0.002**
In-hospital Mortality					
Baseline	0.621	0.027	0.044	0.175	0.015
w/ L_{ind}	**0.641**	**0.029**	0.035	0.128	0.012
w/ L_{coh}	0.625	0.028	0.040	0.166	0.014
CURA	**0.641**	**0.029**	**0.029**	**0.124**	**0.011**

Table 2: Ablation study on **BioGPT** for high-stakes mortality tasks. We isolate the contributions of individual calibration L_{ind} and cohort alignment L_{coh}. Standard deviations are omitted for brevity.

4.4 Ablation Study

We perform ablation studies on BioGPT to isolate the contributions of the two proposed uncertainty components in our objective. Starting from an internal baseline that uses only class-weighted cross-

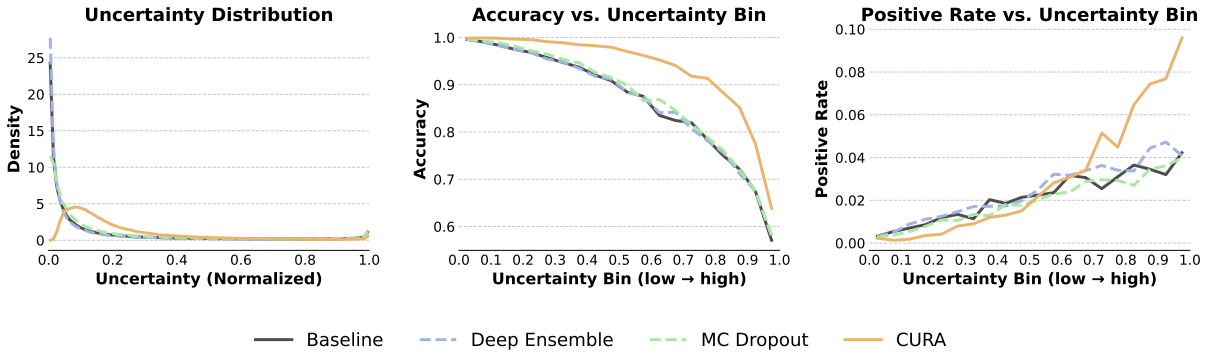

Figure 3: **Left:** Distribution of normalized uncertainty estimates, **Middle:** prediction accuracy within each uncertainty bin at a fixed decision threshold (0.5) on predicted risk, and **Right:** positive rate within each uncertainty bin for the 7-day mortality task with BioGPT.

entropy, we incrementally add (i) the individual calibration term L_{ind}, (ii) the cohort-aware risk alignment term L_{coh}, and (iii) both terms, which together yield the full CURA objective. In the main text we report ablations on two clinically high-stakes tasks, 7-day mortality and in-hospital mortality, and defer the remaining three tasks to Appendix D. As shown in Table 2, across both tasks, adding either L_{ind} or L_{coh} consistently improves calibration metrics such as Brier score, NLL, and AURC, while leaving AUROC and AUPRC essentially unchanged or slightly improved. Combining both terms yields the best overall calibration performance without sacrificing discriminative accuracy. Additionally, we evaluate the sensitivity of CURA to the loss weights λ_{ind} and λ_{coh} on two high-stakes mortality tasks, with detailed results provided in Appendix E.

5 Uncertainty Analysis for Clinical Triage

In this section, we analyze how CURA reshapes uncertainty in deployment-relevant clinical triage, focusing on the BioGPT backbone and the 7-day mortality task where calibration gains are most pronounced. Unless otherwise noted, per-patient uncertainty is measured by the normalized entropy score $u(x)$ from Section 3.2.2, and all results are averaged over five cross-validation folds.

5.1 Uncertainty-driven Risk Stratification

Figure 3 illustrates the interplay between predictive uncertainty, accuracy, and clinical risk. The left panel displays the distribution of uncertainty $u(x)$, showing that the baseline, deep ensembles, and MC dropout all exhibit severe overconfidence, with probability mass heavily concentrated near zero uncertainty. In contrast, CURA produces a dispersed uncertainty distribution, effectively shifting probability mass from the overconfident spike to moderate-to-high uncertainty regions.

The middle panel reports classification accuracy within each uncertainty bin. CURA generally achieves higher bin-wise accuracy than other methods, particularly in the moderate-to-high uncertainty range. Combined with the distribution shift, this indicates that CURA both reduces the number of predictions that the model considers nearly certain and makes each uncertainty level more trustworthy.

Furthermore, the right panel plots the positive rate within each uncertainty bin. All methods show a correlation between uncertainty and risk, but CURA demonstrates the strongest alignment: low-uncertainty bins are dominated by negative cases, whereas high-uncertainty bins are substantially enriched with positive cases. Clinically, this pattern is desirable, since low-uncertainty predictions form a safer pool of patients for automatic triage, whereas high-uncertainty predictions correspond to high-risk, ambiguous cases that should be escalated for doctor review.

Overall, our method mitigates overconfidence and concentrates clinical risk in the high-uncertainty region, yielding uncertainty estimates that are better aligned with risk stratification.

5.2 Selective Deployment and Triage Workload

We evaluate the utility of calibrated uncertainties in supporting selective deployment, a paradigm where the model automates predictions for high-confidence patients while deferring uncertain cases to clinicians. Figure 5 illustrates the AUPRC on the

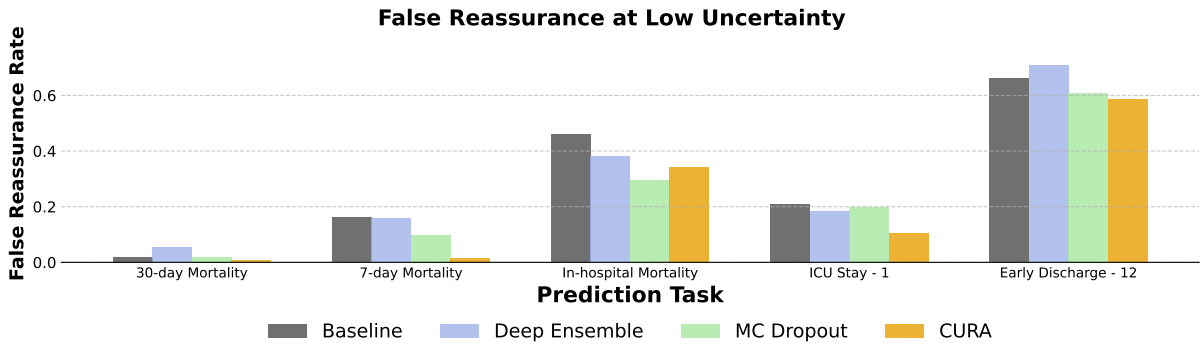

Figure 4: False reassurance rates at low uncertainty across five prediction tasks with BioGPT. Lower values indicate fewer high-risk patients being confidently misclassified as safe.

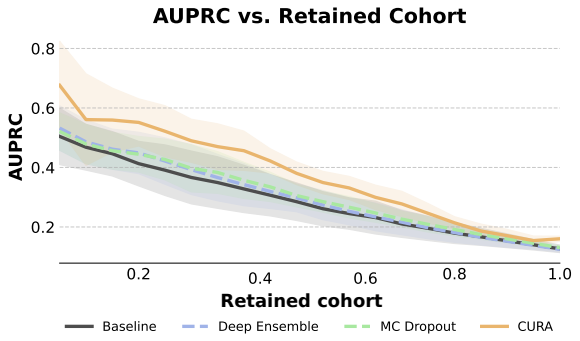

Figure 5: AUPRC on the retained cohort as a function of the retained fraction. Patients are sorted by uncertainty, and only the lowest-uncertainty fraction is retained for prediction.

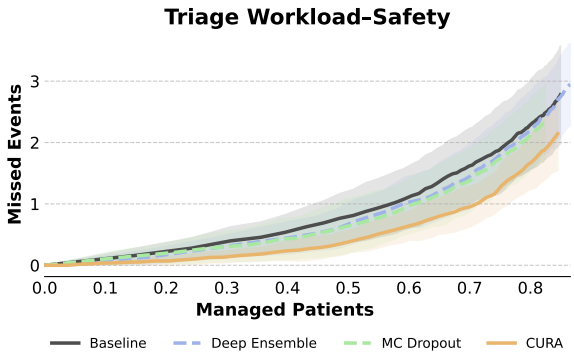

Figure 6: Triage workload–safety trade-off. The curves plot the expected number of missed positive events per 1,000 patients against the fraction of patients automatically managed by the model.

retained cohort, constructed by preserving only the patients with the lowest uncertainty and progressively increasing the retention fraction. CURA consistently outperforms the baselines across the majority of retention levels, indicating better discriminative performance when employed as a decision-support tool for its most confident predictions.

To quantify the clinical implications, Figure 6 depicts the triage workload–safety trade-off. In this analysis, patients are prioritized by uncertainty, with the lowest-uncertainty fraction automatically managed by the model (x-axis); the y-axis reports the expected number of missed positive events per 1,000 patients within this automated group. Across most of the operating range, our method yields fewer missed events than comparison methods for the same level of workload reduction. Equivalently, for a fixed safety threshold, CURA allows a larger proportion of patients to be screened automatically. These results suggest that CURA effectively enhances automation efficiency in high-stakes mortality prediction without compromising patient safety.

5.3 Reducing False Reassurance

We consider a safety metric termed the *false reassurance rate*. Specifically, we define a safe region in which both model uncertainty $u(x) < \tau$ and predicted risk $\bar{p}(x) < \tau$, with $\tau = 0.1$. Additional results using $\tau = 0.05$ and $\tau = 0.15$ are provided in Appendix H, demonstrating consistent trends. The false reassurance rate is the fraction of truly positive patients whose predictions fall into this safe region, capturing a critical clinical failure mode: high-risk patients who are confidently misclassified as low-risk. A detailed definition is provided in Appendix G.

Figure 4 shows the false reassurance rate across all five clinical tasks. CURA generally achieves the lowest rate in four out of five tasks, with substantial reductions in 7-day and 30-day mortality prediction. It remains competitive on in-hospital mortality, where MC Dropout achieves a marginally lower value. These results indicate that our method reduces the fraction of high-risk patients who might

be given misleadingly reassuring predictions.

6 Conclusion

In this work, we propose CURA, a bi-level uncertainty calibration framework that calibrates clinical LMs by jointly optimizing individual predictive uncertainty and cohort-level local consistency. By grounding risk estimates in patient-embedding neighborhoods, CURA mitigates overconfidence, especially on ambiguous cases. Extensive experiments on MIMIC-IV demonstrate that our method consistently improves calibration metrics while largely maintaining competitive discriminative performance. Moreover, our method substantially lowers false reassurance rates and yields uncertainty estimates that better support clinical triage.

Limitations

Despite the promising results of our approach, it has several limitations. Our current framework relies on discriminative LMs providing scalar risk scores without textual explanations, and we do not evaluate closed-source API models, largely because the MIMIC-IV data use agreement prohibits sharing patient data with third-party platforms. Our experiments are restricted to the retrospective analysis of single-center clinical notes. Future work will explore integrating CURA with generative models to offer interpretable reasoning and extending the framework to multi-modal, longitudinal settings and broader patient populations.

Ethical Considerations

Digital health datasets such as MIMIC reflect systemic health disparities experienced by minority populations, which can be inadvertently propagated by AI models. Prior work has documented such disparities in MIMIC-III and other electronic health record datasets in the United States (Röösli et al., 2022; Tripathi et al., 2020; Wang et al., 2024b). Because we build on clinical language models pretrained on MIMIC-III and evaluate on MIMIC-IV, our findings should be interpreted in light of these fairness concerns, and our framework does not directly mitigate underlying structural biases in the data.

Acknowledgments

This work was supported in part by the Fullgraf Foundation and the AI for Health Institute at Washington University in St. Louis. Charles Alba was partially supported by the National University of Singapore Development Grant and the Danforth Scholarship at Washington University in St. Louis. Claire Najjuuko was partially supported by the NIH Researcher Resilience Training Grant (R25MH118935-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, and 1 others. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

Charles Alba, Bing Xue, Joanna Abraham, Thomas Kannampallil, and Chenyang Lu. 2025. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. *npj Digital Medicine*, 8(1):95.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Ziyi Chen, Mengyuan Zhang, Mustafa Mohammed Ahmed, Yi Guo, Thomas J George, Jiang Bian, and Yonghui Wu. 2025a. Narrative feature or structured feature? a study of large language models to identify cancer patients at risk of heart failure. In *AMIA Annual Symposium Proceedings*, volume 2024, page 242.

Zizhang Chen, Peizhao Li, Xiaomeng Dong, and Pengyu Hong. 2025b. Uncertainty quantification for clinical outcome predictions with (large) language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7512–7523.

Keyu Duan, Zichen Liu, Xin Mao, Tianyu Pang, Changyu Chen, Qiguang Chen, Michael Qizhe Shieh, and Longxu Dou. 2025. Efficient process reward model training via active learning. *arXiv preprint arXiv:2504.10559*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, and 1 others. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025a. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Liangjie Huang, Dawei Li, Huan Liu, and Lu Cheng. 2025b. Beyond accuracy: The role of calibration in self-improving large language models. *arXiv preprint arXiv:2504.02902*.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, and 1 others. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. Mimic-iv (version 2.2).

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Ranganath Krishnan, Piyush Khanna, and Omesh Tickoo. 2024. Enhancing trust in large language models with uncertainty-aware fine-tuning. *arXiv preprint arXiv:2412.02904*.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu 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.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2024. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*.

Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. 2022. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*.

Mengdi Luo, Yang Chen, Yuan Cheng, Na Li, and He Qing. 2022a. Association between hematocrit and the 30-day mortality of patients with sepsis: a retrospective analysis based on the large-scale clinical database mimic-iv. *PLoS one*, 17(3):e0265758.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Luckeciano C Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. 2024. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37:118052–118085.

Mark L Mitchell. 2005. Frozen section diagnosis for axillary sentinel lymph nodes: the first six years. *Modern pathology*, 18(1):58–61.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. 2022. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. *Scientific Data*, 9(1):24.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Yongqi Tong, Sizhe Wang, Dawei Li, Yifan Wang, Simeng Han, Zi Lin, Chengsong Huang, Jiaxin Huang, and Jingbo Shang. 2024. Optimizing language model's reasoning abilities with weak supervision. *arXiv preprint arXiv:2405.04086*.

Sandhya Tripathi, Bradley A Fritz, Mohamed Abdelhack, Michael S Avidan, Yixin Chen, and Christopher R King. 2020. (un) fairness in post-operative complication prediction models. *arXiv preprint arXiv:2011.02036*.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. 2024a. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*.

Sizhe Wang, Yongqi Tong, Hengyuan Zhang, Dawei Li, Xin Zhang, and Tianlong Chen. 2025. Bpo: Towards balanced preference optimization between knowledge breadth and depth in alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8811–8826.

Yifei Wang, Liqin Wang, Zhengyang Zhou, John Laurentiev, Joshua R Lakin, Li Zhou, and Pengyu Hong. 2024b. Assessing fairness in machine learning models: A study of racial bias using matched counterparts in mortality prediction for patients with chronic diseases. *Journal of biomedical informatics*, 156:104677.

Yu Xie and Charles F Manski. 1989. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302.

Ziqi Xu, Jingwen Zhang, Simon Haroutounian, Hanyang Liu, Zihan Cao, Gabrielle Rose Messner, Harutyun B Alaverdyan, Saivee Ahuja, Rahul Koshy, Joel Hanns, and 1 others. 2025. Incorporating uncertainty in predictive models using mobile sensing and clinical data: A case study on persistent post-surgical pain. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):1–33.

Bing Xue, York Jiao, Thomas Kannampallil, Bradley Fritz, Christopher King, Joanna Abraham, Michael Avidan, and Chenyang Lu. 2022. Perioperative predictions with interpretable latent representation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4268–4278.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

Muhammad Imran Zulfiqar, Ayesha Khalid, Liming Chen, and Sahraoui Dhelim. 2025. Uncertainty-aware neighbor calibration for positive and unlabeled learning in large machine learning models. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

A MIMIC-IV Dataset and Task Definitions

In this section, we describe the MIMIC-IV dataset used in our experiments and provide detailed definitions of the five clinical risk prediction tasks.

A.1 Task Definitions

7-day mortality We consider the cohort of adult inpatients with at least one discharge summary. This task targets imminent post-discharge deterioration. The label is assigned as positive if the patient dies within 7 days after hospital discharge, with a positive rate of 0.88%.

30-day mortality This task utilizes the same cohort as the 7-day mortality task but extends the prediction window to assess medium-term prognosis. The label is positive if the patient dies within 30 days after hospital discharge, and negative otherwise. The positive rate is 2.87%.

Early discharge within 12 hours This task identifies admissions suitable for rapid triage or those with short, uncomplicated stays. The cohort is restricted to patients who were discharged alive (excluding in-hospital mortality). The label is positive if the discharge occurs within 12 hours of admission, and negative if the stay exceeds 12 hours. The positive rate is 0.74%.

In-hospital mortality For this task, the label is positive if the patient dies before hospital discharge and negative otherwise. It corresponds to the standard in-hospital mortality flag in MIMIC and is widely used as a benchmark for clinical risk prediction models. The positive rate is 1.71%.

ICU stay longer than one day We consider encounters that include an ICU admission record. The label is positive if the patient spends more than 24 hours in the ICU during the index stay (summed across all ICU transfers within the same admission), and negative otherwise. This task reflects early critical-care resource utilization and is relevant for planning ICU staffing and bed management. The positive rate is 2.50%.

After preprocessing, the two post-discharge mortality tasks (7-day and 30-day) contain approximately 323K encounters. For the remaining three tasks (early discharge within 12 hours, in-hospital mortality, and ICU stay longer than one day), the processed cohort consists of approximately 142K encounters. Note that for the early discharge task, we exclude encounters resulting in in-hospital death to decouple mortality outcomes from resource utilization metrics.

A.2 Data Governance and De-identification

All experiments use the MIMIC-IV database of electronic health records and clinical notes (Johnson et al., 2023b), which is released under the PhysioNet Credentialed Health Data License 1.5.0 (Johnson et al., 2023a). Access to the dataset requires completion of human-subjects protection training and acceptance of the MIMIC data use agreement. A credentialed author on our team fulfilled these requirements and accessed the data on behalf of the project, and all analyses complied with the data use agreement.

The MIMIC-IV notes were de-identified by the dataset curators in accordance with the HIPAA Safe Harbor provision. In particular, multiple categories of protected health information (e.g., personal names, locations, serial numbers, and ages) were removed or replaced with surrogate identifiers, and calendar dates were systematically perturbed to reduce the risk of re-identification (Johnson et al., 2023b). We treat the resulting corpus as de-identified and did not attempt to recover any original identifiers.

A.3 Clinical Notes and Preprocessing

We focus on discharge summaries from MIMIC-IV, which are written by clinicians after hospital discharge to summarize the patient's hospital course, treatments, procedures, and follow-up plans. Because these notes were heavily de-identified to meet HIPAA's Safe Harbor standard, many entities identified as protected health information were removed and replaced with placeholder tokens such as "___". As a result, the raw text often contains long sequences of underscores and other artifacts.

To obtain a cleaner corpus for modeling, we adapt preprocessing pipelines from prior work on clinical language models (Huang et al., 2019; Alsentzer et al., 2019; Alba et al., 2025). Specifically, we remove excessive placeholder tokens (e.g., long runs of underscores), normalize whitespace, and discard non-printable or unrecognized characters. We do not attempt to reconstruct protected health information or to enrich the notes with external structured data.

A.4 Exclusion Criteria

We include adult inpatient encounters with at least one discharge summary and with sufficient follow-up information to define the outcome labels described below. Patients without any recorded dis-

charge note are excluded. We further remove encounters whose discharge note is timestamped after or during the recorded time of in-hospital death. Manual inspection of a small sample of such cases indicates that these notes primarily document the death itself rather than the clinical course preceding it. Excluding these encounters avoids potential label leakage from the text into the prediction target and yields a more realistic evaluation setting.

B Additional Experimental Results on Different Models

In this section, we provide a detailed performance analysis using two additional clinical language model backbones: BioClinicalBERT and ClinicalBERT. These experiments aim to verify the generalizability of the CURA framework across different architectures.

For BioClinicalBERT, the results in Table 3 exhibit trends that are highly consistent with our main findings on BioGPT. CURA maintains competitive discriminative performance across all five tasks, often matching or slightly surpassing the baseline. More importantly, our method demonstrates substantial advantages in calibration. While Deep Ensembles and MC Dropout struggle to improve calibration metrics over the baseline on the 7-day mortality task (with Brier scores remaining around 0.070), CURA significantly reduces the Brier score to 0.056 and the NLL from 0.246 to 0.187. This indicates that our method effectively handles uncertainty even in architectures different from that used in our primary analysis.

The results for ClinicalBERT, shown in Table 4, further reinforce these observations. In the 30-day mortality prediction task, while existing uncertainty quantification methods provide marginal gains in discrimination, their impact on calibration is limited. In contrast, our method reduces the NLL from 0.334 (Baseline) to 0.289, achieving the best calibration performance among all compared methods. Furthermore, in the Early Discharge task, CURA achieves by far the largest reduction in AURC (from 0.082 to 0.025), whereas the other methods provide only modest improvements or even degrade performance.

Collectively, these additional experiments confirm that the benefits of our bi-level uncertainty fine-tuning objective are model-agnostic. The framework consistently reshapes the probability distributions to align with risks, delivering reliable uncertainty estimates without compromising the discriminative power of the underlying clinical language models.

C Implementation Details

C.1 Clinical LM Fine-Tuning

For all experiments, we train a separate task-specific model for each outcome and adopt a five-fold cross-validation scheme. We train on the training split for 5 epochs with a learning rate of 1×10^{-5}, for all models and all tasks without early stopping.

C.2 Uncertainty Fine-Tuning and Baselines

For all methods, we adopt the same multi-head classifier architecture described in Section 3.2.1 on top of the frozen embeddings, and use the same hyperparameters during training. We train for up to 50 epochs with a learning rate of 1×10^{-4}, applying early stopping based on validation negative log-likelihood (NLL). The methods differ only in the training objective and in how uncertainty scores are derived from the heads, so that performance differences primarily reflect the uncertainty objective rather than model capacity.

Uncertainty Fine-Tuning In Section 3.2, we attach a 32-head classifier on top of the frozen embeddings and optimize it with the full uncertainty-aware objective L_{total} in Section 3.2.3. Each head is implemented as a lightweight MLP that shares the same input features but maintains independent parameters.

We set the individual calibration weight to $\lambda_{ind} = 0.5$ and the cohort-alignment weight to $\lambda_{coh} = 0.01$ for all experiments. When computing the neighborhood risk $q(x)$ in Equation 4, we retrieve K nearest neighbors using cosine distance. Specifically, we use $K = 200$ nearest neighbors for the 7-day and 30-day mortality tasks, and $K = 100$ neighbors for the remaining three tasks, reflecting their smaller cohort sizes. Neighborhoods $\mathcal{N}_K(e_i)$ are pre-computed once using frozen embeddings to avoid computational overhead during uncertainty training. To prevent information leakage, the query sample itself is excluded from its neighbor set $\mathcal{N}_K(e_i)$.

MC Dropout We train a single multi-head prediction network per task on top of the frozen embeddings. At inference time, we keep dropout active with a probability of $p = 0.5$ and perform $T = 10$

stochastic forward passes. The final prediction is obtained by averaging the predictive probabilities across these passes for each example.

Deep Ensembles We train an ensemble of $M = 5$ independently initialized classifier heads for each task, each using the same multi-head architecture as in Section 3.2.1. At test time, we average the output probabilities from all ensemble members.

C.3 Uncertainty in Analyses

Across all methods, we compute a scalar uncertainty $u(x)$ as the normalized binary entropy of the final predicted positive-class probability. In practice, the binary entropy is computed using the natural logarithm, so the normalization constant $H_{\max}$ in Section 3.2.2 corresponds to the maximum entropy $\log 2$ for a Bernoulli distribution. Using this unified definition allows us to compare triage behavior and risk–coverage curves under a common uncertainty scale, rather than confounding the results with method-specific uncertainty proxies.

D Additional Ablation Results

In Section 4.4, we analyzed the contributions of our uncertainty-aware components on two high-stakes tasks. Here, we extend this analysis to the remaining three outcomes on BioGPT: 30-day mortality, ICU stay longer than one day, and early discharge within 12 hours. The detailed results are reported in Table 5.

These results corroborate the findings presented in the main text. We observe that introducing either the individual calibration term L_{ind} or the cohort-aware alignment term L_{coh} consistently reduces calibration error (Brier score, NLL, and AURC) compared to the baseline. Notably, for the 30-day mortality task, L_{ind} alone drives substantial gains in calibration while simultaneously boosting discriminative performance. Across all three tasks, the full CURA objective—combining both terms—achieves the best overall calibration performance. Furthermore, this improvement in reliability does not come at the cost of accuracy; CURA maintains, and in most cases slightly improves, the AUROC and AUPRC scores relative to the baseline.

E Sensitivity to Loss Weights

The relative strengths of the individual- and cohort-level regularizers are controlled by λ_{ind} in Equation 3 and λ_{coh} through $w(x_i)$ in Equation 5. To evaluate the robustness of these choices, we conduct a sensitivity analysis on BioGPT for the two high-stakes mortality tasks: 7-day mortality and in-hospital mortality. We first vary λ_{ind} while fixing $\lambda_{coh} = 0.01$, and then vary λ_{coh} while fixing $\lambda_{ind} = 0.5$. All other training and evaluation settings are identical to the main experiments.

Tables 6 and 7 show that λ_{ind} around 0.5 gives a strong balance between discrimination and calibration on both tasks. Increasing λ_{ind} beyond 0.5 yields only modest additional calibration gains and can slightly hurt AUROC/AUPRC. For λ_{coh}, small positive values in the range 0.01–0.05 consistently improve calibration metrics, whereas larger values offer no clear further benefit. Overall, these results indicate that our default setting ($\lambda_{ind} = 0.5, \lambda_{coh} = 0.01$) lies in a stable region rather than being finely tuned to a single coefficient choice.

F Cohort-Aware Loss as Soft-Label Cross-Entropy

For completeness, we show that the combination of the base risk loss L_{base} and the cohort-aware loss L_{coh} can be equivalently written as a single cross-entropy with a cohort-informed soft label.

Consider a single example x with binary label $y \in \{0, 1\}$, cohort risk $q(x) \in [0, 1]$, and ensemble-averaged prediction $\bar{p}(x) \in (0, 1)$. Denote the binary cross-entropy between a prediction p and target r by

$$\mathrm{CE}(p, r) = -\big[r \log p + (1 - r) \log(1 - p)\big]. \tag{7}$$

The per-example contributions of the base loss and the cohort-aware loss are

$$\begin{aligned} \ell_{\text{base}}(x) &= \mathrm{CE}\big(\bar{p}(x), y\big), \\ \ell_{\text{coh}}(x) &= w(x)\, \mathrm{CE}\big(\bar{p}(x), q(x)\big), \end{aligned} \tag{8}$$

where $w(x) = \lambda_{\text{coh}}\, \hat{H}\big(q(x)\big) \geq 0$. Their sum can be written as

$$\begin{aligned} &\ell_{\text{base}}(x) + \ell_{\text{coh}}(x) = \\ &- \Big[y \log \bar{p}(x) + \big(1 - y\big) \log\big(1 - \bar{p}(x)\big)\Big] \\ &- w(x)\Big[q(x) \log \bar{p}(x) \\ &+ \big(1 - q(x)\big) \log\big(1 - \bar{p}(x)\big)\Big]. \end{aligned} \tag{9}$$

Grouping the coefficients of $\log \bar{p}(x)$ and

Task	Method	AUROC ↑	AUPRC ↑	Brier ↓	NLL ↓	AURC ↓
7-day Mortality	Baseline	0.863 (0.011)	0.080 (0.011)	0.068 (0.009)	0.246 (0.029)	0.030 (0.023)
	Deep Ensemble	0.863 (0.009)	0.081 (0.007)	0.070 (0.010)	0.242 (0.027)	0.025 (0.004)
	MC Dropout	**0.864** (0.009)	0.081 (0.008)	0.070 (0.010)	0.250 (0.024)	0.025 (0.009)
	CURA (Ours)	0.863 (0.011)	**0.082** (0.008)	**0.056** (0.011)	**0.187** (0.023)	**0.010** (0.003)
30-day Mortality	Baseline	**0.876** (0.005)	**0.228** (0.009)	0.101 (0.003)	0.316 (0.006)	0.033 (0.002)
	Deep Ensemble	0.874 (0.005)	0.226 (0.011)	0.097 (0.002)	0.304 (0.003)	0.031 (0.001)
	MC Dropout	**0.876** (0.005)	**0.228** (0.009)	0.101 (0.003)	0.315 (0.007)	0.033 (0.002)
	CURA (Ours)	**0.876** (0.005)	0.226 (0.007)	**0.088** (0.003)	**0.267** (0.005)	**0.021** (0.001)
In-hospital Mortality	Baseline	0.675 (0.015)	**0.033** (0.004)	0.138 (0.005)	0.422 (0.011)	0.076 (0.005)
	Deep Ensemble	0.671 (0.016)	**0.033** (0.004)	0.131 (0.006)	0.404 (0.015)	0.070 (0.006)
	MC Dropout	0.676 (0.015)	**0.033** (0.004)	0.139 (0.006)	0.426 (0.014)	0.077 (0.006)
	CURA (Ours)	**0.678** (0.013)	**0.033** (0.003)	**0.116** (0.007)	**0.348** (0.017)	**0.046** (0.005)
ICU Stay - 1	Baseline	0.624 (0.006)	0.044 (0.003)	0.150 (0.008)	0.470 (0.017)	0.128 (0.009)
	Deep Ensemble	0.621 (0.007)	0.043 (0.004)	0.140 (0.009)	0.445 (0.020)	0.109 (0.006)
	MC Dropout	0.626 (0.008)	**0.045** (0.004)	0.151 (0.009)	0.473 (0.018)	0.129 (0.006)
	CURA (Ours)	**0.630** (0.006)	0.044 (0.003)	**0.123** (0.009)	**0.375** (0.019)	**0.057** (0.006)
Early Discharge - 12	Baseline	0.655 (0.014)	0.066 (0.019)	0.120 (0.020)	0.406 (0.053)	0.095 (0.021)
	Deep Ensemble	0.654 (0.014)	0.066 (0.014)	0.111 (0.023)	0.381 (0.063)	0.082 (0.023)
	MC Dropout	0.655 (0.015)	0.063 (0.016)	0.130 (0.022)	0.431 (0.056)	0.112 (0.027)
	CURA (Ours)	**0.656** (0.016)	**0.068** (0.013)	**0.100** (0.024)	**0.317** (0.064)	**0.042** (0.017)

Table 3: Performance of **BioClinicalBERT** on five clinical risk prediction tasks. Results are reported as mean (standard deviation) across five-fold cross-validation. "ICU Stay - 1" refers to ICU stays longer than one day, and "Early Discharge - 12" refers to early discharge within 12 hours.

Task	Method	AUROC ↑	AUPRC ↑	Brier ↓	NLL ↓	AURC ↓
30-day Mortality	Baseline	**0.876** (0.003)	0.228 (0.013)	0.111 (0.001)	0.334 (0.004)	0.034 (0.001)
	Deep Ensemble	**0.876** (0.003)	**0.230** (0.012)	0.109 (0.001)	0.327 (0.003)	0.034 (0.001)
	MC Dropout	**0.876** (0.003)	0.229 (0.012)	0.111 (0.001)	0.333 (0.003)	0.035 (0.001)
	CURA (Ours)	**0.876** (0.003)	0.228 (0.013)	**0.098** (0.002)	**0.289** (0.002)	**0.024** (0.001)
7-day Mortality	Baseline	0.859 (0.010)	0.080 (0.007)	0.073 (0.005)	0.241 (0.016)	0.018 (0.003)
	Deep Ensemble	0.859 (0.008)	**0.083** (0.009)	0.071 (0.004)	0.234 (0.010)	0.017 (0.002)
	MC Dropout	0.859 (0.008)	0.082 (0.009)	0.073 (0.005)	0.241 (0.015)	0.019 (0.003)
	CURA (Ours)	**0.860** (0.008)	0.077 (0.006)	**0.064** (0.003)	**0.196** (0.007)	**0.010** (0.001)
In-hospital Mortality	Baseline	**0.685** (0.012)	**0.034** (0.002)	0.169 (0.005)	0.478 (0.010)	0.097 (0.005)
	Deep Ensemble	0.684 (0.012)	**0.034** (0.002)	0.165 (0.007)	0.466 (0.016)	0.093 (0.008)
	MC Dropout	**0.685** (0.012)	**0.034** (0.002)	0.170 (0.005)	0.480 (0.010)	0.098 (0.005)
	CURA (Ours)	0.684 (0.012)	**0.034** (0.002)	**0.150** (0.006)	**0.419** (0.013)	**0.071** (0.006)
ICU Stay - 1	Baseline	0.645 (0.013)	0.045 (0.004)	0.188 (0.006)	0.551 (0.013)	0.157 (0.013)
	Deep Ensemble	0.643 (0.011)	**0.046** (0.004)	0.182 (0.004)	0.535 (0.012)	0.144 (0.011)
	MC Dropout	0.645 (0.013)	0.045 (0.003)	0.190 (0.006)	0.558 (0.015)	0.162 (0.016)
	CURA (Ours)	**0.647** (0.013)	**0.046** (0.004)	**0.170** (0.010)	**0.483** (0.023)	**0.109** (0.013)
Early Discharge - 12	Baseline	0.668 (0.014)	0.067 (0.017)	0.103 (0.007)	0.353 (0.018)	0.082 (0.015)
	Deep Ensemble	0.666 (0.014)	**0.071** (0.016)	0.092 (0.006)	0.320 (0.015)	0.062 (0.006)
	MC Dropout	0.667 (0.015)	0.068 (0.016)	0.107 (0.007)	0.364 (0.018)	0.085 (0.013)
	CURA (Ours)	**0.670** (0.014)	0.056 (0.012)	**0.082** (0.005)	**0.262** (0.017)	**0.025** (0.004)

Table 4: Performance of **ClinicalBERT** on five clinical risk prediction tasks. Results are reported as mean (standard deviation) across five-fold cross-validation. "ICU Stay - 1" refers to ICU stays longer than one day, and "Early Discharge - 12" refers to early discharge within 12 hours.

Method	AUROC	AUPRC	Brier	NLL	AURC
30-day Mortality					
Baseline	0.881	0.270	0.064	0.231	0.024
w/ L_{ind}	**0.890**	**0.280**	0.045	0.148	0.012
w/ L_{coh}	0.881	0.268	0.053	0.212	0.023
CURA	**0.890**	**0.280**	**0.038**	**0.146**	**0.009**
ICU Stay - 1					
Baseline	0.579	0.035	0.105	0.350	0.064
w/ L_{ind}	**0.584**	**0.035**	0.086	0.295	0.049
w/ L_{coh}	0.579	0.035	0.094	0.330	0.063
CURA	**0.584**	**0.035**	**0.085**	**0.288**	**0.040**
Early Discharge - 12					
Baseline	0.587	0.027	0.018	0.085	0.007
w/ L_{ind}	**0.594**	**0.031**	0.013	0.067	**0.006**
w/ L_{coh}	0.587	0.028	0.016	0.076	**0.006**
CURA	**0.594**	**0.031**	**0.010**	**0.056**	**0.006**

Table 5: Full ablation study results on **BioGPT** for the remaining clinical tasks. We isolate the contributions of individual calibration L_{ind} and cohort alignment L_{coh}. Standard deviations are omitted for brevity.

$\log\big(1-\bar{p}(x)\big)$ yields

$$\begin{aligned}&\ell_{\text{base}}(x)+\ell_{\text{coh}}(x)=\\&-\Big[\big(y+w(x)q(x)\big)\log\bar{p}(x)\\&+\big(1-y+w(x)(1-q(x))\big)\log\big(1-\bar{p}(x)\big)\Big].\end{aligned}\tag{10}$$

Define the mixing weight

$$\gamma(x)=\frac{w(x)}{1+w(x)}\in[0,1),\tag{11}$$

and the cohort-aware soft label

$$\begin{aligned}t(x)&=(1-\gamma(x))\,y+\gamma(x)\,q(x)\\&=\frac{y+w(x)q(x)}{1+w(x)}.\end{aligned}\tag{12}$$

Using $1-t(x)=\frac{1-y+w(x)(1-q(x))}{1+w(x)}$, we can rewrite the sum as

$$\ell_{\text{base}}(x)+\ell_{\text{coh}}(x)=(1+w(x))\,\mathrm{CE}\big(\bar{p}(x),t(x)\big).\tag{13}$$

Taking expectations over (x,y) gives

$$L_{\text{base}}+L_{\text{coh}}=\mathbb{E}_{(x,y)}\big[(1+w(x))\,\mathrm{CE}\big(\bar{p}(x),t(x)\big)\big],\tag{14}$$

showing that our objective is equivalent to optimizing a cross-entropy loss with a cohort-informed soft label $t(x)$ and a sample-dependent weight $1+w(x)$. This interpretation highlights L_{coh} as a form of neighborhood-informed label smoothing that focuses more strongly on ambiguous patient cohorts.

G Definition of False Reassurance Rate

In the diagnostic testing literature, the *false reassurance rate* is commonly used to quantify the risk of missed disease among patients who receive a negative test result. Classical definitions take the false reassurance rate to be the complement of the negative predictive value, that is

$$\text{FRR}=1-\text{NPV}=\frac{\text{false negatives}}{\text{false negatives}+\text{true negatives}},$$

representing the proportion of diseased individuals among all test-negative patients (Mitchell, 2005). This notion captures an important failure mode in screening programmes: patients who are incorrectly reassured by a negative result.

In our setting, models output both a predicted risk score $p(x)$ and an uncertainty score $u(x)$ for each patient x. We are specifically concerned with the subset of high-risk patients who are assigned to a region that appears *both* low-risk and highly certain. To formalize this, we define a safe region using thresholds τ:

$$\mathcal{S}=\{x: u(x)<\tau,\ p(x)<\tau\}.$$

We then define the task-specific *False Reassurance Rate (FRR)* as

$$\text{FRR}=\frac{\big|\{x: y=1,\ x\in\mathcal{S}\}\big|}{\big|\{x: y=1\}\big|},\tag{15}$$

where $y=1$ denotes that the patient experiences the adverse outcome, and $|\cdot|$ denotes set cardinality. Intuitively, FRR measures the fraction of truly positive patients whose predictions fall into the low-uncertainty, low-risk region $\mathcal{S}$, i.e., high-risk patients who are confidently misclassified as safe.

Our definition is conceptually related to the classical false reassurance rate, but differs in two aspects. First, the denominator conditions on all positive cases rather than all test-negative cases, reflecting the perspective of "among high-risk patients, how many are given misleading reassurance?". Second, by restricting to the safe region $\mathcal{S}$, we only count those false negatives for which the model is simultaneously low-risk and low-uncertainty, which are the most concerning from a triage standpoint. In the experiments reported in Section 5.3, we set $\tau=0.1$ for all tasks.

Method / Setting	AUROC↑	AUPRC↑	Brier↓	NLL↓	AURC↓
Baseline	0.852	0.127	0.032	0.120	0.008
Vary λ_{ind} ($\lambda_{coh} = 0.01$)					
$\lambda_{ind} = 0.05$	0.859	0.127	0.032	0.121	0.007
$\lambda_{ind} = 0.10$	0.887	0.150	0.030	0.127	0.006
$\lambda_{ind} = 0.20$	0.891	0.157	0.025	0.110	0.004
$\lambda_{ind} = 0.50$	0.892	0.160	0.015	0.075	0.002
$\lambda_{ind} = 1.00$	0.890	0.159	0.012	0.075	0.002
Vary λ_{coh} ($\lambda_{ind} = 0.5$)					
$\lambda_{coh} = 0.005$	0.892	0.160	0.016	0.081	0.003
$\lambda_{coh} = 0.010$	0.892	0.160	0.015	0.075	0.002
$\lambda_{coh} = 0.050$	0.893	0.161	0.013	0.072	0.001
$\lambda_{coh} = 0.100$	0.892	0.160	0.015	0.076	0.002
$\lambda_{coh} = 0.500$	0.892	0.161	0.016	0.078	0.003

Table 6: Sensitivity of CURA to λ_{ind} and λ_{coh} on BioGPT for 7-day mortality. All other training and evaluation settings are identical to the main experiment. Standard deviations are omitted for brevity.

Method / Setting	AUROC↑	AUPRC↑	Brier↓	NLL↓	AURC↓
Baseline	0.621	0.027	0.044	0.175	0.015
Vary λ_{ind} ($\lambda_{coh} = 0.01$)					
$\lambda_{ind} = 0.05$	0.630	0.028	0.042	0.169	0.014
$\lambda_{ind} = 0.10$	0.638	0.029	0.042	0.164	0.013
$\lambda_{ind} = 0.20$	0.642	0.029	0.038	0.150	0.012
$\lambda_{ind} = 0.50$	0.641	0.029	0.029	0.124	0.011
$\lambda_{ind} = 1.00$	0.639	0.028	0.023	0.113	0.010
Vary λ_{coh} ($\lambda_{ind} = 0.5$)					
$\lambda_{coh} = 0.005$	0.642	0.029	0.031	0.129	0.011
$\lambda_{coh} = 0.010$	0.641	0.029	0.029	0.124	0.011
$\lambda_{coh} = 0.050$	0.642	0.029	0.026	0.116	0.009
$\lambda_{coh} = 0.100$	0.641	0.029	0.031	0.125	0.012
$\lambda_{coh} = 0.500$	0.641	0.029	0.033	0.128	0.013

Table 7: Sensitivity of CURA to λ_{ind} and λ_{coh} on BioGPT for in-hospital mortality. All other training and evaluation settings are identical to the main experiment. Standard deviations are omitted for brevity.

H Sensitivity Analysis of False Reassurance Thresholds

In Section 5.3, we presented False Reassurance Rates (FRR) using a fixed safety threshold of $\tau = 0.1$ for both uncertainty and predicted risk. To evaluate the robustness of CURA against this hyperparameter choice, we report FRR results using stricter ($\tau = 0.05$) and looser ($\tau = 0.15$) thresholds in Figure 7 and Figure 8, respectively. Results across both alternative thresholds exhibit patterns consistent with the main-text findings at $\tau = 0.1$. Tightening the threshold to $\tau = 0.05$ constrains the safe region and lowers FRR for all methods; however, CURA generally maintains the lowest or near-lowest false reassurance rates on the acute mortality and early-discharge tasks, while remaining competitive with top baselines on in-hospital mortality and ICU stay prediction. Conversely, relaxing the threshold to $\tau = 0.15$ universally increases FRR, yet CURA continues to yield fewer confidently misclassified high-risk patients than alternative approaches, demonstrating particularly distinct gains on the 7-day mortality, 30-day mortality, and early-discharge tasks. Taken together with Figure 4, these sensitivity analyses indicate that CURA's reduction in false reassurance is not dependent on a finely tuned choice of τ, and its safety benefits persist under both stricter and more permissive operating points of the BioGPT backbone.

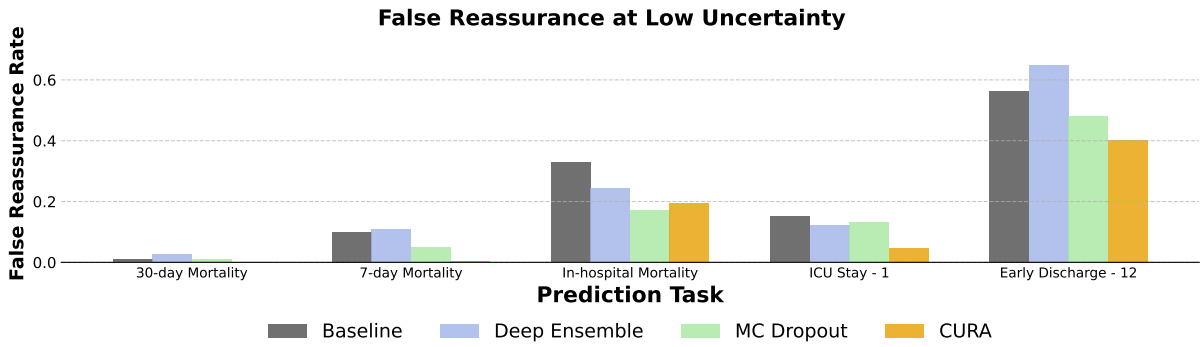

Figure 7: False reassurance rates ($\tau = 0.05$) at low uncertainty across five prediction tasks with BioGPT. Lower values indicate fewer high-risk patients being confidently misclassified as safe.

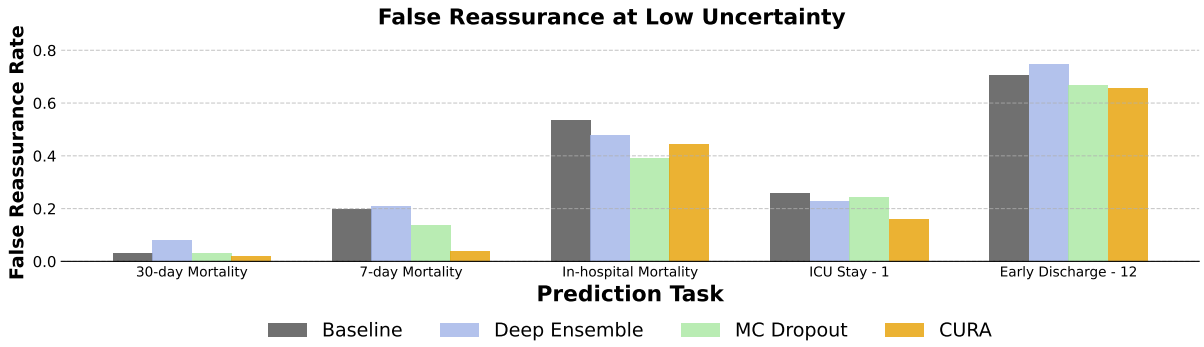

Figure 8: False reassurance rates ($\tau = 0.15$) at low uncertainty across five prediction tasks with BioGPT. Lower values indicate fewer high-risk patients being confidently misclassified as safe.

I Runtime Comparison

To assess the practical overhead of CURA, we compare the total uncertainty fine-tuning runtime of different methods under identical hardware, classifier architecture, and early-stopping settings on the 7-day mortality task with the BioGPT backbone. All methods use the same extracted embeddings and the same multi-head classifier, so the comparison isolates the overhead introduced by the uncertainty mechanism itself. For CURA, the K-nearest-neighbor cohorts are pre-computed once at the beginning of each run, and this pre-computation time is included in the reported GPU hours.

As shown in Table 8, CURA has runtime comparable to, and in our setting slightly shorter than, the single-model baseline, while remaining substantially more efficient than Deep Ensembles. This result supports our claim that CURA functions as a lightweight plug-in objective: it adds two loss terms computed from existing probabilities and pre-computed neighborhoods, without requiring multiple independently trained backbones or repeated stochastic inference, resonating with broader trends toward lightweight or data-efficient post-training objectives, including uncertainty-aware fine-tuning, weak supervision, active-learning-based reward modeling, and balanced preference optimization (Krishnan et al., 2024; Tong et al., 2024; Wang et al., 2025; Duan et al., 2025).

Method	Time / GPU hours
Baseline	1.03
Deep Ensemble	4.43
MC Dropout	1.08
CURA (Ours)	**0.94**

Table 8: Runtime comparison of uncertainty fine-tuning methods on the 7-day mortality task with the BioGPT backbone. Values denote total GPU hours on a single NVIDIA A40 GPU under identical hardware, classifier architecture, and early-stopping settings, aggregated over five-fold cross-validation. For CURA, the one-time pre-computation of K-nearest-neighbor cohorts is included in the reported runtime.

J Computing Infrastructure and Resources

J.1 Model Sizes

We experiment with three backbone language models. BERT-based models such as ClinicalBERT and BioClinicalBERT contain approximately 110 million parameters, while the GPT-style BioGPT backbone contains about 347 million parameters.

J.2 Hardware and Compute Environment

Most experiments were conducted on A40 GPUs with 48 GB of VRAM provided by Research Infrastructure Services compute1 cluster, which is a secure service approved for ePHI, protected, or confidential data.

K Use of AI Assistants

AI assistants were used only for grammar and syntax revision of the manuscript. All experiments, analyses, and substantive writing were conceived and produced by the authors without AI-generated content.