TRIAGE: DIALECTICAL LLM REASONING FOR EXPLAINABLE RISK PREDICTION ON IRREGULARLY SAMPLED MEDICAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical early warning systems built on irregularly sampled medical time series (ISMTS) from electronic health records must deliver continuous risk scores for patient triage as well as interpretable rationales that clinicians can verify. Large language models (LLMs) are uniquely positioned for both, deriving risk from their output probabilities and rationales from their medical knowledge. However, we find that conventional LLM reasoning collapses graded risk into overconfident predictions and thereby undermines the cross-patient comparability on which triage depends. We refer to this failure mode as *risk polarization* and identify two underlying behaviors: early commitment to a single outcome, and one-sided reasoning that focuses only on the evidence for that outcome. To address this, we propose **TRIAGE**, a framework that trains an LLM to reason dialectically over competing clinical outcomes by eliciting outcome-specific rationales. This dialectical formulation mitigates risk polarization, enabling a single LLM to jointly provide explicit clinical rationales and risk scores comparable across patients. Across five ISMTS benchmarks, TRIAGE improves mean AUPRC by 17.0% and reduces mean calibration error by 82.8% relative to the competitive LLM-based baseline, while surpassing the strongest ISMTS baseline by 3.5% in mean AUPRC.[1]

1 INTRODUCTION

Predicting adverse clinical events from electronic health records (EHRs) is central to early warning systems in critical care. In EHRs, clinical observations are typically recorded sparsely at irregular intervals with substantial missingness and are thus commonly treated as irregularly sampled medical time series (ISMTS) (Sun et al., 2020). An effective clinical decision support system for ISMTS must satisfy two requirements in tandem that existing systems rarely provide together: (1) a *well-calibrated, cross-patient comparable risk score* for triage and resource allocation (Van Calster et al., 2019; Foroutan et al., 2026), and (2) a *natural language explanation* grounded in reasoning over clinical knowledge that clinicians can evaluate (Hou & Wang, 2025).

Existing approaches to this task focus solely on risk prediction. Specialized deep learning models for ISMTS (Che et al., 2018; Horn et al., 2020; Luo et al., 2024) achieve strong predictive performance through dedicated architectures tailored to irregular sampling and inter-variable dependencies, but produce opaque predictions with no accompanying rationale. Post-hoc explainability methods (Lundberg & Lee, 2017; Crabbé & Van Der Schaar, 2021) can be applied to identify contributing variables or time points, but their explanations are limited to feature-level or temporal importance, rather than the higher-level abstractions that clinicians reason over.

Large language models (LLMs) are uniquely positioned to bridge this gap as their output distributions provide a continuous signal for risk estimation, while their learned medical knowledge enables them to articulate complex reasoning in natural language. Nevertheless, existing LLM-based approaches still do not satisfy both requirements within a single prediction process (Table 1). One line of work (Belyaeva et al., 2023; Liao et al., 2025) extracts a continuous risk score from the model's implicit answer-token probability, thereby treating the LLM as a classifier without any natural language reasoning. A second line of work (Jiang et al., 2025a; Langer et al., 2026) elicits reasoning

[1]The code is available at https://anonymous.4open.science/r/Anonymous_ICLR-F259/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of ISMTS risk prediction approaches. *Natural language reasoning* covers any rationales or summaries in text format. *Score explained by reasoning*: the reasoning explains the continuous risk score rather than being its input or by-product.

Method	Continuous risk score	Natural language reasoning	Score explained by reasoning
Specialized ISMTS architectures (Che et al., 2018; Tipirneni & Reddy, 2022; Li et al., 2026)	✓	×	×
LLMs as predictors (Jin et al., 2024; Belyaeva et al., 2023; Liao et al., 2025)	✓	×	×
LLMs as reasoners, hard label (Jiang et al., 2025a; Liu et al., 2025a; Langer et al., 2026)	×	✓	×
LLMs as reasoners, composite pipeline (Lee et al., 2025; Ji et al., 2026; Yu et al., 2026)	✓	✓	×
TRIAGE (Ours)	✓	✓	✓

alongside the prediction but optimizes solely for the hard label, yielding discrete risk estimates that are not comparable across patients. Hybrid pipelines pair LLM-generated interpretations with a neural predictor (Lee et al., 2025; Ji et al., 2026), but the risk score and the rationale for the prediction originate from separate components and are therefore not inherently coupled.

We first scrutinize why current LLMs struggle to meet both requirements: comparable risk scores and natural language explanations. Our preliminary study uncovers a *risk polarization problem*: when reasoning is elicited before risk scoring, the scores shift toward the extremes. We further identify two recurring reasoning behaviors that contribute to this polarization. First, rationales frequently commit to a single outcome *in advance* of the final prediction (*e.g.*, *"Therefore, this patient is likely to die. Answer:"*), making the resulting prediction nearly deterministic given the preceding biased text. Second, rationales are typically *one-sided*, presenting evidence for the committed outcome while omitting evidence for alternatives. This issue is particularly problematic for ISMTS, where trajectories can contain signals of both deterioration and stabilization (Arnold et al., 2009; Hernandez et al., 2014; Jang et al., 2025). Consequently, conventional LLM reasoning is fundamentally insufficient to provide both clinically grounded explanations and cross-patient comparable risk scores.

Motivated by these observations, we propose **T**ime Series **R**easoning by **I**nspecting **A**lternative Outcomes for **G**rounded Risk **E**stimation (**TRIAGE**), a framework that mitigates risk polarization by grounding risk estimation in dialectical reasoning over alternative outcomes. TRIAGE generates a dedicated rationale for each candidate outcome, requiring the LLM to articulate patient-specific evidence for each possible alternative before making the final prediction. It then derives the final risk from the LLM's implicit probability, without any intervening verdicts. This design preserves a graded risk estimate while making the evidence considered for each outcome explicit. TRIAGE learns this behavior through a two-stage training pipeline: *Dialectical Reasoning Supervision*, which teaches the LLM to generate outcome-specific rationales, followed by *On-policy Self-refinement*, which uses reinforcement learning to further improve risk discrimination and calibration.

Experiments on five ISMTS benchmarks spanning mortality and sepsis prediction show that TRIAGE improves both risk discrimination and calibration, achieving an improvement of **17.0%** in mean AUPRC and a reduction of **82.8%** in mean expected calibration error (ECE) compared with the strongest LLM-based baseline, while also surpassing the strongest specialized ISMTS baseline by 3.5% in mean AUPRC. We further verify the medical correctness and clinical helpfulness of TRIAGE's dialectical rationales through case studies and human evaluation by clinical experts.

2 Related Work

Specialized neural architectures for ISMTS. Prior work has developed specialized architectures for irregularly sampled time series, including RNN-based models (Che et al., 2018), continuous-time models (Chen et al., 2018; Rubanova et al., 2019; Biloš et al., 2021; Gao et al., 2026), set-based encoders (Horn et al., 2020), interpolation-based models (Shukla & Marlin, 2019), attention-based models (Shukla & Marlin, 2021; Zhang et al., 2023), transformer-based models (Tipirneni & Reddy, 2022; Li et al., 2023), and graph-based models (Zhang et al., 2022; Luo et al., 2024; 2025;

Li et al., 2026). While these methods advance representation learning and predictive accuracy on ISMTS, they do not provide any natural language explanations, which clinicians have found more understandable and reasonable than common explainable AI (XAI) methods (Hou & Wang, 2025).

LLMs as predictors without reasoning. One line of LLM-based approaches uses LLMs directly as risk predictors. Some use the LLM as a *pretrained backbone*, attaching a separate task head that maps its hidden states to a continuous prediction (Jin et al., 2024; Pan et al., 2024; Hu et al., 2025; Zhang et al., 2025a). Others adopt *answer-token risk scoring*, treating the LLM as a classifier and reading a risk score from its answer-token probabilities, without an additional head (Belyaeva et al., 2023; Liao et al., 2025). Nevertheless, neither yields a rationale as part of the prediction.

LLMs as reasoners. Another line generates natural language reasoning, but in neither of its forms does the reasoning explain a continuous risk score. Hard-label methods produce a rationale followed by a discrete class label (Jiang et al., 2025a; Liu et al., 2025a; Langer et al., 2026), leaving no continuous score with which to compare patients. Composite pipelines pair the generated text with a separate predictor, embedding LLM-generated rationales or summaries as features for a downstream predictor (Nguyen et al., 2024; Ji et al., 2026), combining a discrete LLM prediction with a time-series encoder (Lee et al., 2025; Jiang et al., 2025b), or fusing the LLM's hidden states with a separate encoder while generating explanations in parallel (Yu et al., 2026). Since a downstream predictor or fusion module produces the final score, the text is an input to or a by-product of that score rather than its explanation.

Existing work thus uses LLMs as risk scorers without generated reasoning, reasoners for hard-label prediction, or auxiliary components in composite predictors. TRIAGE instead unifies reasoning and risk prediction in a single LLM, generating outcome-specific rationales and deriving a continuous risk score from the implicit probability conditioned on those rationales. This autoregressive formulation directly couples explanation and prediction.

3 Problem Setup: Clinical Risk Prediction

We formulate clinical risk prediction from ISMTS as a supervised classification problem. Let $\mathcal{D} = \{(\mathbf{s}_i, \mathbf{z}_i, y_i)\}_{i=1}^{N}$ denote a cohort of N patients. The temporal record $\mathbf{s}_i$ is a multivariate ISMTS represented as $\mathbf{s}_i = \{(t_{i,k}, v_{i,k}, m_{i,k})\}_{k=1}^{L_i}$, where L_i is the total number of observations and each tuple specifies the time $t_{i,k}$, value $v_{i,k}$, and variable index $m_{i,k} \in \{1, \ldots, D\}$ of the k-th measurement among D clinical variables. The static descriptor $\mathbf{z}_i \in \mathbb{R}^{d_s}$ encodes time-invariant patient attributes such as age and gender, and $y_i \in \{0, \ldots, C-1\}$ is the clinical outcome label among C classes.

Our objective is to learn a predictor that jointly outputs a patient-specific class distribution $\mathbf{p}_i = (p_{i,0}, \ldots, p_{i,C-1})$ and a natural language rationale r_i grounded in the patient's record, where $p_{i,c} \geq 0$ is the probability that patient i belongs to class c. While a discrete prediction can be obtained as $\hat{y}_i = \arg\max_c p_{i,c}$, our primary quantity of interest is the probability vector, which yields a continuous risk estimate over the candidate outcomes that is comparable across patients.

4 Preliminary Study: Diagnosing Risk Polarization

In this section, we identify why standard LLM reasoning fails at clinical risk prediction. A natural approach to explainable risk prediction is to have the LLM reason over the patient record before making a prediction, ensuring that the risk score is grounded in explicit clinical evidence. Through a preliminary study, we show that this approach instead collapses the score distribution into a degenerate extreme. We term this failure **risk polarization** and identify its two underlying causes. Full details of the preliminary study are provided in Appendix B.

Setup. We study in-hospital mortality prediction on MIMIC-III (Johnson et al., 2016) over 3,228 patients with gpt-oss-120b (Agarwal et al., 2025) and Qwen3-14B (Yang et al., 2025), reporting metrics for both discrimination (AUROC, AUPRC) and calibration (ECE, BS). The *risk score* refers to the model's *implicit* probability: the softmax probability assigned to the positive-class answer token derived from the output logits, rather than any verbalized number. This implicit probability provides a finer-grained and directly trainable risk signal in our setting (Appendix B.2).

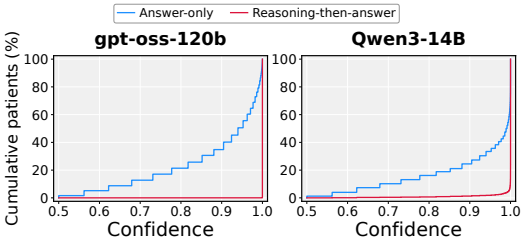

Figure 1: Cumulative distribution of confidence on MIMIC-III in-hospital mortality prediction.

Reasoning causes risk polarization. We compare *reasoning-then-answer*, where the model generates a rationale before predicting (*e.g.*, `[Rationale]`→`Yes/No`), against an *answer-only* baseline (*e.g.*, `Yes/No` without reasoning). Under *answer-only* prediction, confidence, the larger of the two outcome probabilities $\max_c p_{i,c}$, spreads over $[0.5, 1]$ for both models (Figure 1). Under *reasoning-then-answer*, however, it concentrates near 1; for gpt-oss-120b, even the lowest confidence across all patients is 0.9998. The risk score then reduces to a near-binary label and gives no meaningful basis for ranking patients.

Reasoning often pre-commits to an outcome. Under the *reasoning-then-answer* paradigm, regardless of the preceding reasoning quality, rationales tend to terminate in an explicit verdict immediately before the answer position (*e.g.*, *"Therefore, this patient is likely to die. Answer:"*). To quantify this tendency, we use gpt-oss-120b as the LLM judge to classify whether each rationale closes with such a verdict (judging instruction in Appendix H), and find that this pattern appears in 71.7% of rationales from gpt-oss-120b and 44.1% from Qwen3-14B. Once such a verdict appears in the prefix, it dominates the conditioning context. Although the clinical evidence remains in the context, the answer token is overwhelmingly steered by the nearby verdict, concentrating the predicted-class probability on the outcome already stated in the rationale. The rationale thus acts as a self-imposed hard prior that absorbs residual uncertainty before the label is produced.

Reasoning is typically one-sided. Beyond the final verdict, the reasoning itself lacks balanced deliberation. Inspecting traces reveals that the model exhibits *confirmation bias*, committing to one outcome and citing only supportive evidence while disregarding countervailing signals (see Appendix B.4). This conflicts with clinical trajectories, where indicators of deterioration and stabilization may coexist (Arnold et al., 2009; Hernandez et al., 2014; Jang et al., 2025). To test whether this tendency is systematic, we append a minimal two-sided instruction: *"Weigh the evidence for survival and for in-hospital death, then decide."* This improves AUPRC and reduces ECE and Brier score while maintaining comparable AUROC (Table 2); one-sided reasoning is therefore a systematic limitation, motivating dialectical deliberation as a better inductive bias.

Table 2: Results on MIMIC-III in-hospital mortality prediction, with risk estimated as the fraction of 10 sampled responses predicting death. AUROC and AUPRC are in %.

Method	AUROC ↑	AUPRC ↑	ECE ↓	BS ↓
gpt-oss-120b				
Standard	**76.0**	28.8	0.236	0.211
Two-sided	75.5	**30.3**	**0.205**	**0.185**
Qwen3-14B				
Standard	69.3	19.0	0.583	0.519
Two-sided	**70.8**	**20.1**	**0.575**	**0.507**

Takeaways. Our preliminary study thus motivates two design principles: **(P1)** *Reasoning should be structured, avoiding any explicit outcome verdict before risk estimation.* **(P2)** *Reasoning should inspect each candidate outcome separately, identifying the clinical evidence that supports it.*

5 TRIAGE

Guided by the design principles from Section 4, we propose **T**ime Series **R**easoning by **I**nspecting **A**lternative Outcomes for **G**rounded Risk **E**stimation (**TRIAGE**; see Figure 2). TRIAGE introduces two components: (1) *alternative outcome inspection*, which addresses **P2** by generating a dedicated rationale for each candidate outcome, and (2) *verdict-free risk estimation*, which addresses **P1** by terminating the reasoning chain directly with the outcome token, so that no verdict or summary follows the rationales, and deriving a graded risk from the model's implicit distribution at that position (Section 5.1). To learn this reasoning behavior, we introduce a two-stage training pipeline of TRIAGE: (1) *Dialectical Reasoning Supervision* instills the structured reasoning behavior using supervised fine-tuning (SFT) on synthesized outcome-conditioned rationales, and (2) *On-policy Self-refinement* refines the model through reinforcement learning (RL) on its own samples to improve both discrimination and calibration (Section 5.2).

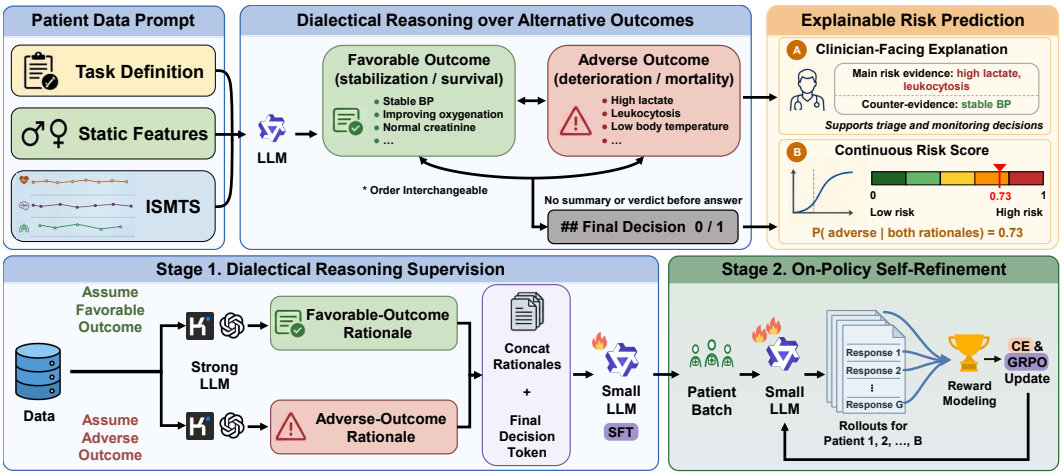

Figure 2: Overview of TRIAGE. **(Top)** Given irregular patient records represented as language inputs, TRIAGE inspects alternative outcomes dialectically to jointly produce a patient risk score and clinician-facing explanations. **(Bottom)** The training pipeline consists of dialectical reasoning supervision followed by on-policy self-refinement.

5.1 Dialectical Reasoning over Alternative Outcomes

Input representation. Following set-based encoding for irregular time series (Horn et al., 2020), we serialize only the observed measurements in a variable-centric format rather than representing the full observation matrix. The input prompt x_i for patient i concatenates a structured task definition $\mathcal{P}$, textualized time-invariant attributes $\mathbf{t}_{z_i}$ (*e.g.*, age, gender), and temporally ordered observations $\mathbf{t}_{s_i}$ (see Figure 9 for an example):

$$x_i = \texttt{concat}(\mathcal{P}, \mathbf{t}_{z_i}, \mathbf{t}_{s_i}). \tag{1}$$

Alternative outcome inspection (P2). As shown in Section 4, LLM reasoning tends to align with a single outcome direction, citing only supportive evidence while disregarding evidence for alternative outcomes. This one-sided tendency is especially problematic for ISMTS, where patient trajectories frequently exhibit concurrent indicators of both deterioration and stabilization (Arnold et al., 2009; Hernandez et al., 2014; Jang et al., 2025).

To address this, TRIAGE generates a dialectical reasoning chain that examines each candidate outcome separately before the final prediction, as illustrated in the upper part of Figure 2. For a binary prediction task with candidate outcomes $\mathcal{Y} = \{y^+, y^-\}$, the model produces a dedicated rationale r_{y_k} for each $y_k \in \mathcal{Y}$, articulating the evidence in the patient's observations that supports that outcome:

$$\texttt{chain} = [r_{y_1}, r_{y_2}, \hat{y}], \quad \{y_1, y_2\} = \{y^-, y^+\}, \tag{2}$$

where r_{y_k} is the rationale for outcome y_k, and $\hat{y}$ is the final predicted outcome token. The order of the two rationales is interchangeable. This structure realizes **P2** by requiring the model to consider both sides of the clinical evidence before reaching a decision, surfacing signals that one-sided reasoning would discard.

Verdict-free risk estimation (P1). Even with dialectical rationales, a reasoning chain that concludes with a verdict before $\hat{y}$ reintroduces the risk polarization that motivates **P1**. To preserve a graded risk signal, TRIAGE structures the reasoning chain to end with a "## Final Decision" header followed directly by a single outcome token, without any intermediate summary or verdict. As detailed in Section 5.2, we instill this structure through SFT on chains whose headers delimit each rationale and the final decision. We map each outcome $y_k \in \mathcal{Y}$ to a designated token (*e.g.*, "0" for negative, "1" for positive) and extract the corresponding logit ℓ_k at this fixed position. The risk score is then computed as $P(y_k \mid x) = \exp(\ell_k) / \sum_{k'} \exp(\ell_{k'})$. Because no verdict precedes this position, the distribution at the outcome token stays graded, and the resulting risk estimates are continuous and comparable across patients rather than near-deterministic.

5.2 Two-Stage Training Pipeline

Dialectical Reasoning Synthesis. As illustrated in the lower-left part of Figure 2, we elicit rationales from a strong LLM *separately* for each candidate outcome: the LLM is prompted to assume that a given outcome holds and to identify only the patient features that support it (prompts are provided in Appendix H). This process must adhere to two constraints. First, the LLM must not reference any alternative outcome, ensuring that each rationale remains strictly outcome-specific rather than contrastive. Second, the LLM must not fabricate unobserved evidence. In cases where no supporting evidence exists, the rationale is left blank. These constraints yield concise, feature-grounded rationales while discouraging hallucinated or contrastive content. The resulting rationales are concatenated with the ground-truth answer token to form reasoning traces as in Equation 2.

Stage 1: Dialectical Reasoning Supervision. Using the synthesized reasoning traces above, we fine-tune the target LLM to learn the dialectical reasoning procedure: generating a dedicated rationale for each candidate outcome followed by the ground-truth decision token. Because the ordering of the outcome-specific rationales is interchangeable, we augment each training instance with both orderings to avoid associating a particular outcome with a fixed position in the reasoning chain. The resulting model initializes the on-policy self-refinement in Stage 2.

Stage 2: On-policy Self-refinement. After SFT on off-policy dialectical reasoning traces built from a strong model, the model can generate high-quality rationales. However, its self-generated rationales may differ in distribution from the SFT targets produced by the strong model, leaving it susceptible to exposure bias when conditioning on its own prefixes at inference (Agarwal et al., 2024).

To address this limitation, we further post-train the model with Group Relative Policy Optimization (GRPO; Shao et al., 2024). The cross-entropy objective is applied only to the final decision token to supervise the implicit predictive probability, whereas the preceding reasoning tokens are optimized by a GRPO objective modified following Dr. GRPO (Liu et al., 2025c) and DAPO (Yu et al., 2025):[2]

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) + \lambda \mathcal{L}_{\text{CE}}(\theta), \tag{3}$$

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\{r_j\}_{j=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{L_{\text{r}}}\sum_{j=1}^{G}\sum_{\tau\in\mathcal{T}_j^{\text{r}}}\min\Big(\rho_{j,\tau}\hat{A}_j, \text{clip}(\rho_{j,\tau}, 1-\epsilon_{\text{l}}, 1+\epsilon_{\text{h}})\,\hat{A}_j\Big)\right], \tag{4}$$

$$\text{where}\quad L_{\text{r}} = \sum_{j=1}^{G}|\mathcal{T}_j^{\text{r}}|, \qquad \rho_{j,\tau} = \frac{\pi_\theta(r_{j,\tau} \mid x, r_{j,<\tau})}{\pi_{\theta_{\text{old}}}(r_{j,\tau} \mid x, r_{j,<\tau})}, \qquad \hat{A}_j = R_j - \text{mean}(\{R_k\}_{k=1}^{G}),$$

and $\mathcal{T}_j^{\text{r}}$ denotes the reasoning-token positions excluding the final decision token.

While intra-sample rewards provide a learning signal, they fail to explicitly encourage cross-patient discrimination. To ensure that risk scores are comparable across patients, we define rewards at the batch level. Inspired by Yuan et al. (2021), we define the reward for the j-th response of patient i as

$$R_{i,j} = -\begin{cases} \dfrac{1}{|\mathcal{B}^-|}\displaystyle\sum_{i'\in\mathcal{B}^-}[m-(\sigma_{i,j}-\bar{\sigma}_{i'})]_+^2 & \text{if } y_i = 1, \\ \dfrac{1}{|\mathcal{B}^+|}\displaystyle\sum_{i'\in\mathcal{B}^+}[m-(\bar{\sigma}_{i'}-\sigma_{i,j})]_+^2 & \text{if } y_i = 0, \end{cases} \tag{5}$$

where $\sigma_{i,j}$ is the log-odds of the j-th response $r_{i,j}$ for patient i, $\bar{\sigma}_i = \frac{1}{G}\sum_{j=1}^{G}\sigma_{i,j}$, and $[z]_+ = \max\{z, 0\}$. Here, m is the margin, and $\mathcal{B}^-$ and $\mathcal{B}^+$ denote the indices of negative and positive patients in the batch, respectively. The reward pushes the log-odds of each response at least m above (for positive patients) or below (for negative patients) the mean log-odds of each opposite-class patient in the batch. This further strengthens the cross-patient comparability and calibration that on-policy training with a sample-level reward already improves (ablation in Appendix G.1).

6 Experiments

6.1 Experimental Setup

Datasets and metrics. We conduct experiments on five ISMTS benchmarks, P12 (Silva et al., 2012), P19 (Reyna et al., 2019), MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al.,

[2] We apply a KL penalty term $\beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ during training, which is omitted here for brevity.

Table 3: Discrimination performance on clinical risk prediction benchmarks. Specialized ISMTS models (gray) cannot explain their predictions and are shown for reference only; **bold** and underline mark the best and second-best results among natural-language-explainable methods. For clarity, we report only the top-performing and recent ISMTS baselines; full results are provided in Appendix D.

Method	P12		P19		MIMIC-III		MIMIC-IV		eICU	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Specialized ISMTS Models (no natural-language explanation)										
GRU-D	87.1±0.9	56.3±2.3	89.7±0.9	56.6±2.9	84.7±0.2	49.0±0.6	87.5±0.5	59.2±1.4	84.4±0.7	56.4±2.2
STraTS	87.3±1.5	57.4±3.9	88.7±1.4	50.4±2.3	85.3±0.5	48.3±0.6	88.6±0.4	61.3±1.1	85.3±0.7	57.4±2.6
KEDGN	87.8±1.1	58.3±3.2	88.2±0.5	53.5±3.0	84.7±0.4	48.5±0.9	87.9±0.5	60.1±1.2	85.3±0.6	56.4±1.4
Hi-Patch	87.3±0.8	57.0±3.7	88.7±1.6	52.1±4.3	84.6±0.6	46.2±0.3	87.7±0.3	57.7±0.6	85.0±0.8	56.5±2.5
GSNF	84.7±1.3	50.7±2.4	86.2±2.0	46.2±3.2	83.6±0.4	40.5±0.7	86.0±0.5	54.6±1.6	83.4±0.7	54.4±1.1
GARLIC	86.6±0.8	55.3±2.2	88.8±0.6	52.4±1.8	85.5±0.3	50.5±1.5	87.2±0.5	57.9±1.8	85.2±1.0	58.2±1.5
Zero-shot LLMs										
GPT-5.1	83.9±1.0	49.3±0.5	72.2±2.9	9.3±0.9	–	–	–	–	–	–
gpt-oss-120b	81.2±0.7	43.4±3.0	64.2±2.1	6.7±0.6	76.6±0.6	31.0±1.0	70.8±0.5	27.9±1.0	79.5±1.2	47.4±1.0
Qwen3-4B	69.7±1.9	26.7±0.8	56.6±1.7	5.0±0.6	69.3±0.5	20.2±0.2	65.6±0.7	20.9±0.8	71.2±1.4	31.4±0.9
LLM-based Prediction Methods										
Record2Vec	76.9±1.3	38.2±2.9	79.5±1.6	33.4±3.8	73.6±0.1	31.9±0.1	79.5±1.2	43.6±1.9	75.9±0.5	42.5±1.1
TimeCAP	85.2±1.0	52.8±3.9	85.4±1.3	33.7±2.3	84.2±0.1	46.9±0.2	87.3±0.3	57.5±0.5	83.8±1.0	54.6±3.2
TRIAGE$_{\text{SFT}}$	86.9±1.0	56.4±1.9	88.9±1.0	52.2±3.2	86.4±0.3	51.4±0.9	87.5±0.2	59.7±0.8	84.2±1.3	56.7±0.9
TRIAGE	**87.3**±1.2	**59.0**±3.4	**89.3**±1.0	**53.8**±1.4	**86.7**±0.2	**54.1**±0.5	**87.8**±0.5	**61.6**±2.1	**85.2**±0.5	**58.7**±1.4

Table 4: Calibration performance on clinical risk prediction benchmarks. ECE and Brier score (BS); lower is better. The best results are indicated in **bold**.

Method	P12		P19		MIMIC-III		MIMIC-IV		eICU	
	ECE	BS	ECE	BS	ECE	BS	ECE	BS	ECE	BS
GPT-5.1	0.085	0.103	0.161	0.084	–	–	–	–	–	–
gpt-oss-120b	0.263	0.185	0.321	0.187	0.244	0.174	0.172	0.161	0.196	0.165
Qwen3-4B	0.427	0.309	0.679	0.527	0.609	0.484	0.355	0.262	0.490	0.382
Record2Vec	0.298	0.212	0.292	0.137	0.311	0.207	0.269	0.187	0.256	0.195
TimeCAP	0.245	0.190	0.215	0.113	0.230	0.155	0.180	0.138	0.225	0.173
TRIAGE$_{\text{SFT}}$	0.190	0.149	0.149	0.091	0.214	0.153	0.174	0.142	0.172	0.156
TRIAGE	**0.044**	**0.089**	**0.039**	**0.035**	**0.032**	**0.077**	**0.035**	**0.080**	**0.038**	**0.102**

2023), and eICU (Pollard et al., 2018).[3] P19 involves predicting sepsis onset within 6 hours, while the other four target in-hospital mortality. For discrimination, we report the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPRC), with AUPRC as the primary metric given the severe class imbalance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). For calibration, we report the expected calibration error (ECE) and the Brier score (BS). We additionally evaluate robustness under limited information, a practically relevant setting in real-world clinical environments (Appendix E), and report the inference cost of every method (Appendix A).

Baselines. We compare against representative ISMTS baselines, namely GRU-D (Che et al., 2018), mTAND (Shukla & Marlin, 2021), SeFT (Horn et al., 2020), Raindrop (Zhang et al., 2022), STraTS (Tipirneni & Reddy, 2022), ViTST (Li et al., 2023), KEDGN (Luo et al., 2024), Hi-Patch (Luo et al., 2025), GSNF (Gao et al., 2026), and GARLIC (Li et al., 2026). We further compare with LLM-based methods, TimeCAP (Lee et al., 2025) and Record2Vec (Ji et al., 2026), as well as GPT-5.1[4] (OpenAI, 2025), gpt-oss-120b (Agarwal et al., 2025), and Qwen3-4B (Yang et al., 2025) under zero-shot prompting (scored by verbalized probabilities; Appendix B.3).

[3] MIMIC-III, MIMIC-IV, and eICU are distributed under PhysioNet's credentialed Data Use Agreement, which prohibits sending the data to third-party LLM services (https://physionet.org/news/post/llm-responsible-use/).

[4] GPT-5.1 is not evaluated on MIMIC-III, MIMIC-IV, or eICU due to the data-use restriction (footnote 3).

Implementation details. To collect training data, we use GPT-5.1 (OpenAI, 2025) for P12 and P19, and, in accordance with data-use restrictions, locally served Kimi K2 Thinking (Team et al., 2025) for MIMIC-III, MIMIC-IV, and eICU. Since the datasets are highly imbalanced, we apply oversampling following prior work (Zhang et al., 2022; Luo et al., 2024), collecting multiple distinct rationales per minority sample. We apply the two-stage training pipeline to Qwen3-4B-Base (Yang et al., 2025) (see Appendix G.2 for backbone ablations). Full details are in Appendix C.

6.2 Main Results

Discrimination (Table 3). Zero-shot LLMs rank at the bottom. Larger LLMs do better, but even GPT-5.1 remains far from the specialized ISMTS models: its AUPRC trails GRU-D by 7.0 points on P12 and by 47.3 points on P19. This indicates that general-purpose reasoning alone does not transfer well to clinical risk prediction on ISMTS. With SFT alone, TRIAGE achieves a mean AUPRC of 55.3, which is 12.6% higher than TimeCAP (49.1), the strongest LLM-based prediction method, and within 0.3 points of the best ISMTS baselines, GRU-D (55.5), KEDGN (55.4), and STraTS (55.0). TRIAGE thus matches specialized architectures while still producing natural language rationales. RL adds a further gain. TRIAGE reaches a mean AUPRC of 57.4, a 3.9% relative improvement over its SFT counterpart and 3.5% over GRU-D, with the best AUPRC on four of the five benchmarks and the second best on the remaining one. Mean AUROC follows the same pattern with smaller margins (87.3 for TRIAGE vs. 87.0 for STraTS, the best baseline).

Calibration (Table 4). TimeCAP and Record2Vec remain poorly calibrated, on par with or worse than zero-shot gpt-oss-120b (mean ECE of 0.219 and 0.285 vs. 0.239), and TRIAGE with SFT alone improves on them only modestly (mean ECE 0.180, BS 0.138). RL closes this gap, reducing the mean ECE from 0.180 to 0.038 and the mean BS from 0.138 to 0.077, relative reductions of 79% and 45%. As a result, TRIAGE achieves the lowest ECE and Brier score on every benchmark.

7 Analysis of TRIAGE Reasoning

We analyze the reasoning of TRIAGE from three perspectives:[5] (i) the effect of dialectical reasoning on risk estimation (Section 7.1), (ii) the clinical correctness of the resulting rationales (Section 7.2), and (iii) their faithfulness to the predictions (Section 7.3). Appendix F.1 complements (i) and (ii) with an expert evaluation by clinicians on rationale consistency, correctness, and helpfulness.

7.1 Effectiveness of Dialectical Reasoning

To isolate the effect of reasoning structure, Table 5 compares the SFT checkpoint of TRIAGE before RL refinement with zero-shot inference and two alternative SFT variants. The *answer-only* variant removes reasoning entirely and directly predicts the label, treating the LLM as a classifier. The *one-sided rationale* variant uses only the rationales corresponding to the ground-truth outcomes from our collected data. Detailed prompts are in Appendix H.

Answer-only SFT substantially outperforms zero-shot inference (Table 5) but provides no clinical justification. Adding *one-sided rationales* introduces clinical reasoning but inherits the risk polarization problem identified in Section 4, where the answer-token probability collapses toward 1.0. Even after averaging predictions over 10 sampled outputs at $10\times$ inference cost, it underperforms *answer-only* SFT on both metrics, indicating that one-sided supervision harms risk estimation. In contrast, TRIAGE's dialectical reasoning achieves the best AUROC (86.9%) and AUPRC (56.4%) while retaining clinically grounded explanations. Beyond discrimination, clinicians rate the dialectical rationales as more helpful than one-sided ones ($+0.62$, $p < 0.001$), with correctness unchanged (Appendix F.1).

Table 5: Ablation on the reasoning structure.

Method	AUROC	AUPRC
Zero-shot inference	69.7±1.9	26.7±0.8
SFT baselines		
Answer-only	86.4±1.5	53.4±2.2
One-sided rationale†	83.8±1.0	43.1±3.6
TRIAGE$_{\text{SFT}}$	**86.9**±1.0	**56.4**±1.9

† Risk estimated by averaging over 10 sampled responses.

[5] Unless otherwise noted, we report all analyses on P12 with Qwen3-4B-Base.

7.2 Case Studies of TRIAGE Reasoning

We qualitatively examine whether TRIAGE provides clinically coherent rationales that preserve temporal context. Figures 10 and 11 present two illustrative cases. For comparison, we apply Integrated Gradients (IG; Sundararajan et al., 2017) post-hoc to STraTS, a strong ISMTS baseline on P12, and use GPT-5.1 to interpret the resulting textualized attributions. The following interpretations are grounded in the medical literature and reviewed by two clinicians.

For the survivor (Figure 10), TRIAGE treats the consistently maximal GCS of 15 as favorable and HCO_3 of 19 mEq/L as adverse evidence, consistent with *GCS 15 indicating intact consciousness* (Munakomi et al., 2026) and *low HCO_3 potentially indicating metabolic acidosis* (Castro et al., 2024). In contrast, IG attributions for STraTS carry the opposite sign for both. TRIAGE also tracks the urine output trajectory from 5–30 mL/h to 300–400 mL/h, consistent with *rising urine output indicating renal recovery* (Khwaja, 2012), whereas STraTS surfaces only isolated late measurements.

For the non-survivor (Figure 11), TRIAGE identifies severely elevated early potassium and elevated troponin T as adverse evidence (Simon & Rout, 2025; Babuin et al., 2008), while recognizing the later decline in potassium to 4.3 mEq/L as countervailing evidence. Neither appears among the high-attribution IG features of STraTS, which instead place WBC 23.3 and early BUN 98–99 mg/dL on the favorable side, despite *the former representing marked leukocytosis* (Mank et al., 2026) and *the latter being markedly elevated and associated with ICU mortality* (Arihan et al., 2018).

We further evaluate the rationales on consistency, correctness, and helpfulness through human expert evaluation (Appendix F.1), with an LLM-as-a-judge evaluation as a scalable *proxy* (Appendix F.2).

7.3 Faithfulness Analysis of TRIAGE Reasoning

We test whether the rationales of TRIAGE are faithful, that is, whether the predicted risk depends on the evidence they cite, following the input-removal protocol of Crabbé & Van Der Schaar (2021). For 1,000 patients, we randomly select a fraction of the sentences in the rationale for the predicted outcome, locate the measurements they cite with GPT-5.1, and erase them before re-running the model (*Rationale-based*). The control erases the same number of observed values at random (*Random*), which separates reliance on the cited evidence from mere input sparsity. Agreement with the full-input prediction is nearly unchanged under *Random*, but falls monotonically under *Rationale-based* and drops below chance at a deletion ratio of 100% (Figure 3).

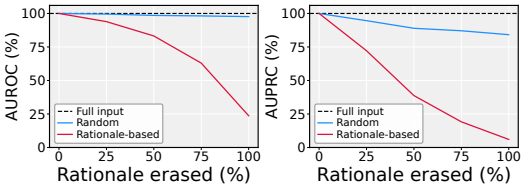

Figure 3: AUROC (left) and AUPRC (right) against the full-input prediction on P12 as a function of the deletion ratio.

We also test whether the rationales stay grounded in the modified input by regenerating them after erasure. At a deletion ratio of 50%, a GPT-5.1 judge, blind to which sentences were targeted, rates how fully each sentence of the original rationale reappears in the regenerated rationale, from retained (1) to absent (5); the prompt is given in Appendix H. Sentences whose evidence was erased average 4.63 (near absent), against 1.86 (near retained) for those left intact. Together with the removal experiment, this indicates that the rationales are both faithful and grounded in the input. A hallucination analysis further shows that only 1.3% of rationales cite a feature or value absent from the record (Appendix F.3).

8 Conclusion

We have introduced TRIAGE, a framework that grounds clinical risk estimation in reasoning over alternative outcomes rather than a single committed trajectory. By inspecting each candidate outcome before deriving the final risk from the LLM's implicit probability, TRIAGE addresses the risk polarization problem that arises when reasoning is elicited before risk scoring. Through dialectical reasoning supervision followed by on-policy self-refinement, TRIAGE outperforms competitive baselines while producing rationales that align with how clinicians weigh patient evidence. More broadly, this work signals a path forward for LLM-based clinical decision support, where predictive performance and grounded explanation are pursued together rather than as competing objectives.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, assisting with translation, interpreting results, and supporting qualitative and thematic data analysis. Specifically, we collected our training data by using a generative AI model, complying with the credentialing requirements as detailed in Section 6. Moreover, we manually implemented our method and utilized AI tools to receive feedback and optimize our code. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments. Additionally, we used generative AI tools for creating scientific figures, editing software code, searching for information, drafting parts of the paper, editing the paper to improve readability, and proposing a title for this paper. We have reviewed all AI-assisted work. We thoroughly verified and tested the LLM-generated feedback and code for correctness and efficiency. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Data governance. MIMIC-III, MIMIC-IV, and eICU are accessed under the PhysioNet Credentialed Data Use Agreement.[3] Accordingly, no patient-level data from these datasets is sent to any third-party LLM service: preprocessing, rationale generation, training, and evaluation (including all zero-shot LLM baselines) are performed locally with self-hosted, open-weight models.

Potential clinical risks. TRIAGE is a research prototype and has not been validated for clinical deployment. Its natural language rationales may contain inaccurate clinical interpretations or reflect biases in the underlying data and models. Over-reliance on these outputs without clinician oversight could lead to harm. The outputs of TRIAGE should not be treated as medical advice or used for autonomous clinical decision-making. Any deployment should follow standard LLM safety practices, with clinical decisions remaining with qualified practitioners.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility and subsequent use by clinicians and practitioners, we release the code for data construction, supervised fine-tuning, and reinforcement learning of TRIAGE, together with our STraTS baseline at an anonymous repository.[1] Additional baseline implementations and final model checkpoints trained on P12 and P19 will be released upon publication. Under the PhysioNet Credentialed Data Use Agreement,[3] models trained on MIMIC-III, MIMIC-IV, and eICU cannot be publicly released. For these datasets, we release the training code together with the preprocessing and rationale-collection instructions so that credentialed users can reproduce our models on the original data. Dataset preprocessing and splits are detailed in Appendix C.1, training procedures and hyperparameters in Appendix C.4, and all prompts in Appendix H. Numerical results may vary slightly due to differences in hardware configurations and the inherent stochasticity of rationale generation and reinforcement learning.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Anthropic. Claude Sonnet 4.5 System Card, September 2025. URL https://www.anthropic.com/claude-sonnet-4-5-system-card.

Okan Arihan, Bernhard Wernly, Michael Lichtenauer, Marcus Franz, Bjoern Kabisch, Johanna Muessig, Maryna Masyuk, Alexander Lauten, Paul Christian Schulze, Uta C Hoppe, et al. Blood

urea nitrogen (bun) is independently associated with mortality in critically ill patients admitted to icu. *PloS one*, 13, 2018.

Ryan C Arnold, Nathan I Shapiro, Alan E Jones, Christa Schorr, Jennifer Pope, Elisabeth Casner, Joseph E Parrillo, R Phillip Dellinger, Stephen Trzeciak, Emergency Medicine Shock Research Network (EMShockNet) Investigators, et al. Multicenter study of early lactate clearance as a determinant of survival in patients with presumed sepsis. *Shock*, 32, 2009.

Luciano Babuin, Vlad C Vasile, Jose A Rio Perez, Jorge R Alegria, High-Seng Chai, Bekele Afessa, and Allan S Jaffe. Elevated cardiac troponin is an independent risk factor for short-and long-term mortality in medical intensive care unit patients. *Critical care medicine*, 36, 2008.

Elizabeth A Baker, Cynthia H Ledford, Louis Fogg, David P Way, and Yoon Soo Park. The idea assessment tool: assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes. *Teaching and learning in medicine*, 27, 2015.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, 2023.

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. *Advances in Neural Information Processing Systems*, 34, 2021.

Danny Castro, Sachin M Patil, Muhammad Zubair, and Michael Keenaghan. Arterial blood gas. In *StatPearls [internet]*. StatPearls Publishing, 2024. URL https://www.ncbi.nlm.nih.gov/books/NBK536919/.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, 2021.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *International Conference on Machine Learning*, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019.

Farid Foroutan, Jessica Weiss, Martin Mayer, Lena Kristin Bache-Mathiesen, Thomas Agoritsas, Per Olav Vandvik, Marisa Signorile, Joie Ensor, Arnav Agarwal, Gordon Guyatt, et al. Discriminating clinical prediction models: how outcome incidence and decision thresholds shape clinical utility and apt auroc values. *Journal of Clinical Epidemiology*, 195, 2026.

Mengzhou Gao, kaiwei wang, and Pengfei Jiao. One-step graph-structured neural flows for irregular multivariate time series classification. In *International Conference on Machine Learning*, 2026.

Google. Gemini 3 flash: frontier intelligence built for speed, December 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bowen Gu, Rishi J Desai, Kueiyu Joshua Lin, and Jie Yang. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

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

Glenn Hernandez, Cecilia Luengo, Alejandro Bruhn, Eduardo Kattan, Gilberto Friedman, Gustavo A Ospina-Tascon, Andrea Fuentealba, Ricardo Castro, Tomas Regueira, Carlos Romero, et al. When to stop septic shock resuscitation: clues from a dynamic perfusion monitoring. *Annals of intensive care*, 4, 2014.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, 2020.

Jun Hou and Lucy Lu Wang. Explainable ai for clinical outcome prediction: A survey of clinician perceptions and preferences. *AMIA Summits on Translational Science Proceedings*, 2025, 2025.

Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. Context-alignment: Activating and enhancing llms capabilities in time series. In *International Conference on Learning Representations*, 2025.

Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-aware integrated gradients for time series explanation. In *International Conference on Machine Learning*, 2025.

Zongliang Ji, Yifei Sun, Andre Carlos Kajdacsy-Balla Amaral, Anna Goldenberg, and Rahul G Krishnan. Can we generate portable representations for clinical time series data using LLMs? In *International Conference on Learning Representations*, 2026.

Pengcheng Jiang, Cao Danica Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations*, 2025a.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. *Advances in Neural Information Processing Systems*, 38, 2025b.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 2023.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in sft-rl post-training: When high sft scores mislead and what to use instead. In *International Conference on Learning Representations*, 2026.

Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120, 2012.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, et al. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *International Conference on Machine Learning*, 2026.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Yanke Li, Ruirui Wang, Manuel Günther, and Diego Paez-Granados. Garlic: Graph attention-based relational learning of multivariate time series in intensive care. In *International Conference on Learning Representations*, 2026.

Zekun Li, Shiyang Li, and Xifeng Yan. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36, 2023.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025.

Haoxin Liu, Chenghao Liu, and B Aditya Prakash. A picture is worth a thousand numbers: Enabling llms reason about time series via visualization. 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)*, 2025a.

Kezhao Liu, Jason Klein Liu, Mingtao Chen, and Yiming Liu. Rethinking kl regularization in rlhf: From value estimation to gradient optimization. *arXiv preprint arXiv:2510.01555*, 2025b.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025c.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Yicheng Luo, Zhen Liu, Linghao Wang, Junhao Zheng, Binquan Wu, and Qianli Ma. Knowledge-empowered dynamic graph network for irregularly sampled medical time series. *Advances in Neural Information Processing Systems*, 37, 2024.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch gnn for irregular multivariate time series. In *International Conference on Machine Learning*, 2025.

Victoria Mank, Waqas Azhar, and Kevin Brown. Leukocytosis. In *StatPearls [internet]*. StatPearls Publishing, 2026. URL https://www.ncbi.nlm.nih.gov/sites/books/NBK560882/.

Sunil Munakomi, Konstantinos Margetis, and Lindsay M Iverson. Glasgow coma scale. In *StatPearls [internet]*. StatPearls Publishing, 2026. URL https://www.ncbi.nlm.nih.gov/sites/books/NBK513298/.

Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Carer-clinical reasoning-enhanced representation for temporal health risk prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

OpenAI. Gpt-5.1 instant and gpt-5.1 thinking system card addendum, November 2025. URL https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf.

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. S^2IP-LLM: Semantic space informed prompt learning with LLM for time series forecasting. In *International Conference on Machine Learning*, 2024.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5, 2018.

Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology*, 2019.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32, 2019.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10, 2015.

Verity Schaye, Louis Miller, David Kudlowitz, Jonathan Chun, Jesse Burk-Rafel, Patrick Cocks, Benedict Guzman, Yindalon Aphinyanaphongs, and Marina Marin. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. *Journal of General Internal Medicine*, 37, 2022.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025.

Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019.

Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, 2012.

Leslie V Simon and Preeti Rout. Hyperkalemia. In *StatPearls [internet]*. StatPearls Publishing, 2025. URL https://www.ncbi.nlm.nih.gov/sites/books/NBK470284/.

Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data. *arXiv preprint arXiv:2010.12493*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16, 2022.

Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, Topic Group 'Evaluating diagnostic tests, and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38, 2025.

Zhengqiu Yu, Yueping Ding, and Xiangrong Liu. PathwayLLM: Explainable clinical trajectory modeling with structured pathways for sepsis prediction. In *International Conference on Machine Learning*, 2026.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Weijia Zhang, Chenlong Yin, Hao Liu, and Hui Xiong. Unleashing the power of pre-trained language models for irregularly sampled time series. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025a.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Appendix

TABLE OF CONTENTS

A Limitations: Computational Cost **17**

B Preliminary Study Details **18**

B.1 Implementation Details 18

B.2 Implicit Probabilities as Risk Scores 18

B.3 Verbalized Probabilities for Zero-Shot Evaluation 18

B.4 One-Sided Reasoning Traces 19

C Experimental Setup Details **19**

C.1 Datasets 19

C.2 Metrics 21

C.3 Baselines 21

C.4 Implementation Details 23

D Full Results on Discrimination **24**

E Robustness under Limited Information **26**

E.1 Leave-Random-Variables-Out 26

E.2 Low-Resource Training 26

F Additional Analysis of TRIAGE Reasoning **27**

F.1 Human Expert Evaluation 27

F.2 Quantitative Analysis 28

F.3 Hallucination Analysis 28

G Further Experiments **29**

G.1 Ablation Study on RL Reward Design 29

G.2 Ablation Study on the Backbone 29

G.3 Inference Direction Analysis 29

H Prompts **30**

Table 6: Inference cost on clinical risk prediction benchmarks. Each entry is inference time per patient in milliseconds, measured on a single B200 GPU. GPT-5.1 is not run locally; its entries are the mean latency of synchronous API calls and are not directly comparable. The last column is the mean over datasets.

Method	P12	P19	MIMIC-III	MIMIC-IV	eICU	Mean
Specialized ISMTS Models (non-explainable in natural language)						
GRU-D	0.157	0.018	0.057	0.300	0.313	0.169
mTAND	0.033	0.013	0.024	0.024	0.018	0.022
SeFT	0.096	0.079	0.091	0.101	0.108	0.095
Raindrop	0.347	0.355	0.368	8.683	0.387	2.028
STraTS	0.471	0.130	0.483	0.304	0.378	0.353
ViTST	3.188	3.178	3.182	3.182	3.187	3.184
KEDGN	0.454	0.058	0.113	10.430	0.423	2.296
Hi-Patch	2.239	0.740	0.590	8.533	2.033	2.827
GSNF	0.239	0.071	0.126	1.377	0.168	0.396
GARLIC	1.250	0.321	0.293	6.744	0.465	1.814
Zero-shot LLMs						
GPT-5.1	12,609	10,878	–	–	–	11,744
gpt-oss-120b	143.0	91.1	132.4	118.4	159.1	128.8
Qwen3-4B	463.9	652.5	468.1	436.4	787.2	561.6
LLM-based Prediction Methods						
Record2Vec	249.5	196.4	264.9	221.5	242.1	234.9
TimeCAP	400.0	332.1	374.7	348.3	345.3	360.1
TRIAGE	223.3	113.6	134.8	170.5	194.7	167.4

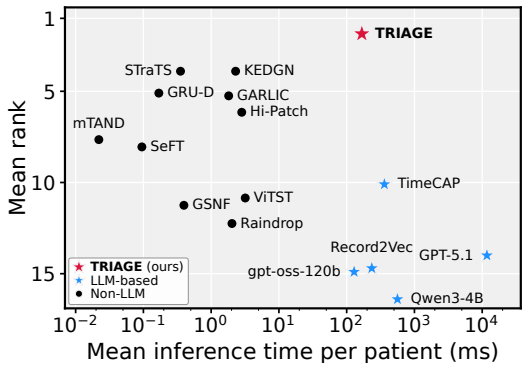

Figure 4: Mean inference time per patient (ms, log scale) versus mean rank over the ten discrimination metrics. Lower is better on both axes.

A Limitations: Computational Cost

The principal limitation of TRIAGE is its inference cost. We measure this cost as the mean wall-clock time per patient on a single B200 GPU. For each method and dataset, we tune the batch size over $\{32, 128, 512, 2048\}$ (and the request concurrency for LLM-based methods) on the validation set and keep the setting with the lowest total time, so each row reflects the throughput-optimal configuration for that method. We then time inference on the test set with this setting and divide the total time by the number of test patients to obtain the per-patient cost in Table 6. GPT-5.1 is queried through its API: its entries are the mean latency of 10 synchronous calls, depend on server-side conditions we do not control, and are listed for reference only.

Figure 4 plots this cost against each method's mean rank over the ten discrimination metrics (AUROC and AUPRC on five datasets) in Table 11. TRIAGE takes 167.4 ms per patient, faster than both LLM-based prediction methods, Record2Vec (234.9 ms) and TimeCAP (360.1 ms), and faster than Qwen3-4B (561.6 ms); among the locally run LLMs, only gpt-oss-120b (128.8 ms) is faster. It also

has a lower mean rank than every LLM-based baseline and every specialized ISMTS model. The specialized ISMTS models are far cheaper, averaging under 4 ms per patient; the fastest, mTAND, takes 0.022 ms, nearly four orders of magnitude less than TRIAGE. We attribute most of this gap to the natural language explanation that TRIAGE generates with each prediction, which the specialized models do not produce. This overhead remains practical for ICU risk scoring, but it becomes a limitation in high-throughput screening, where many patients must be scored quickly, and in hospitals without dedicated GPU servers, since TRIAGE requires serving a 4B-parameter LLM.

B Preliminary Study Details

B.1 Implementation Details

We conduct the preliminary experiment on in-hospital mortality prediction using the MIMIC-III (Johnson et al., 2016) test split ($n = 3{,}228$) with gpt-oss-120b (Agarwal et al., 2025) and Qwen3-14B (Yang et al., 2025). All runs use a temperature of 0.7. For reasoning-then-answer, reasoning is enabled with `reasoning_effort=medium` for gpt-oss-120b and `enable_thinking=True` for Qwen3-14B. We provide the full prompts in Appendix H.

B.2 Implicit Probabilities as Risk Scores

We consider two ways of eliciting probabilities from the LLM. The *verbalized probability* $p^{\text{verb}} \in [0, 1]$ is the numeric estimate in the model's response text, and the *implicit probability* $P(y{=}1 \mid \text{prompt})$ is read off the next-token distribution over the answer tokens. We use the implicit probability as the risk score in the preliminary study (Section 4). This choice is supported by Gu et al. (2024), who report that, across six open-source LLMs and five medical datasets, verbalized probabilities give lower discrimination, precision, and recall than implicit probabilities, with larger gaps for smaller models and imbalanced datasets. Two further properties favor the implicit probability in our setting: it gives finer-grained scores, and it receives a direct training signal.

Granularity. With reasoning disabled, verbalized outputs on MIMIC-III concentrate on a few anchor values (Table 7): seven values cover 75.3% of gpt-oss-120b estimates and 86.5% of Qwen3-14B estimates. Patients who share one of these values receive the same score and cannot be ranked against each other. The implicit probability is continuous in $[0, 1]$; on the same test set it takes 1,375 distinct values for gpt-oss-120b and 137 for Qwen3-14B, and its seven most frequent values cover only 10.5% and 14.0% of outputs.

Table 7: Concentration of risk scores on MIMIC-III in *non-reasoning* mode. Top-k is the share (in %) of outputs falling on the k most frequent values; #Distinct counts unique values produced across all test samples.

Method	Top-1	Top-3	Top-7	#Distinct
gpt-oss-120b				
Verbalized	37.0	55.0	75.3	55
Implicit	2.3	6.0	10.5	1,375
Qwen3-14B				
Verbalized	32.9	59.3	86.5	37
Implicit	2.3	6.5	14.0	137

Training signal. During supervised fine-tuning, the cross-entropy loss on the answer token is the negative log of the implicit probability assigned to the true label, so the binary label supervises it directly for every patient. The verbalized estimate has no direct target for supervised fine-tuning, since the data contain binary outcomes and no ground-truth probabilities, and using the label itself as the target number would teach the model to state only extreme values. Reinforcement learning with a proper-scoring-rule reward can still train it, but the reward scores only the numbers the model samples, and training starts from the coarse, less discriminative estimates described above.

B.3 Verbalized Probabilities for Zero-Shot Evaluation

For zero-shot LLM benchmarking under reasoning, two candidates are available for a graded risk score: (i) an *outcome-based* estimator that approximates the model's underlying probability by averaging binary answers across many sampled responses, and (ii) the *verbalized* probability stated in a single response. The outcome-based estimator surpasses the verbalized score only at around 30

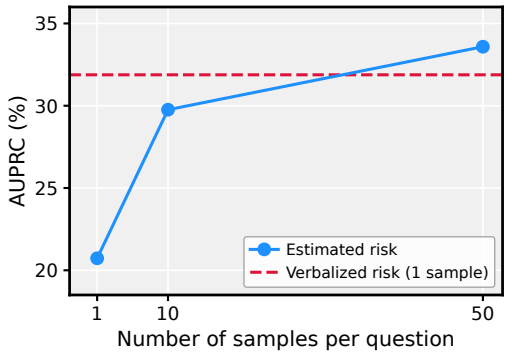

Figure 5: AUPRC of the outcome-based and verbalized risk scores under reasoning, on MIMIC-III with gpt-oss-120b.

sampled responses per question, at a proportional inference cost (Figure 5). The verbalized probability is a clean single-number readout from a single response, providing a far more efficient zero-shot signal. We therefore adopt the verbalized probability whenever reporting zero-shot LLM baselines in subsequent sections.

B.4 One-Sided Reasoning Traces

We observe that LLM reasoning traces typically confirm a single outcome rather than weigh evidence on both sides. MIMIC-III is a credentialed dataset, so we do not reproduce raw patient records here. We instead draw an illustrative case from P12 (Silva et al., 2012), which targets the same task as MIMIC-III (in-hospital mortality prediction from early ICU trajectories) and exhibits the same one-sidedness under our prompting protocol.

To make this behavior easy to inspect, we set `reasoning_effort=None` for gpt-oss-120b and prompt the model with *"Provide your clinical reasoning, then state your final prediction."* This places the entire rationale in the visible response, where the one-sided pattern is straightforward to identify.

Figure 8 shows two reasoning traces out of ten sampled from gpt-oss-120b on the same patient. One trace cites only deterioration signals and predicts death; the other cites only stabilization signals and predicts survival. Each trace is internally coherent, yet neither weighs evidence from the opposite direction.

C Experimental Setup Details

C.1 Datasets

We evaluate our method on five medical time series classification benchmarks: P12 (Silva et al., 2012), P19 (Reyna et al., 2019), MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), and eICU (Pollard et al., 2018). Table 8 summarizes their statistics. All five are binary classification tasks with severe class imbalance. Following Zhang et al. (2022); Li et al. (2023), we use five random 8:1:1 splits for P12, P19, MIMIC-IV, and eICU. For MIMIC-III, we adopt the fixed 7:1.5:1.5 split defined by prior work (Harutyunyan et al., 2019) and run five random seeds. In all cases, we report the mean and standard deviation across the five runs. We describe each dataset below, including our preprocessing refinements over the original releases.

P12. The P12 dataset (Silva et al., 2012) is taken from the PhysioNet 2012 Challenge.[6] Following Zhang et al. (2022); Li et al. (2023), we exclude 12 inappropriate samples, leaving 11,988 ICU patient records. Each record consists of irregularly sampled measurements from 36 sensors collected during the first 48 hours of an ICU stay, together with 5 static demographic features (Age, Gender, Height, ICUType, Weight). The task is binary in-hospital mortality prediction. We adopt the training, validation, and test splits released by Raindrop (Zhang et al., 2022).

[6] https://physionet.org/content/challenge-2012/1.0.0/

Table 8: Statistics of the five medical time series classification datasets. #Avg. obs. denotes the average number of observed measurements per sample. The missing ratio is computed with respect to each patient's actual observation length. Pos. rate denotes the proportion of positive samples.

Dataset	#Samples	#Variables	#Avg. obs.	Missing ratio	Pos. rate
P12	11,988	36	401	84.9%	14.2%
P19	38,738	34	238	80.5%	4.0%
MIMIC-III	21,107	16	431	65.5%	13.2%
MIMIC-IV	26,070	96	341	98.0%	13.4%
eICU	12,312	14	57	65.2%	17.6%

We also correct a parsing misalignment in the original preprocessing, where the five demographic features are extracted by row position rather than by parameter name. When a sensor reading at t=00:00 is interleaved with the demographic rows, the two streams contaminate each other: real sensor values are parsed as demographics, while later static fields such as weight are pushed into the time series. This affects 207 of 11,988 patients (1.7%) and produces implausible values such as height $= 1$ cm.

P19. The P19 dataset (Reyna et al., 2019) is taken from the PhysioNet 2019 Challenge on sepsis prediction.[7] Each sample contains 34 irregularly sampled time series variables observed for up to 60 hours, together with 6 static features (Age, Gender, Unit1, Unit2, HospAdmTime, ICULOS). The task is binary classification of whether sepsis will occur within the next 6 hours. The public release used by prior work (Zhang et al., 2022; Li et al., 2023; Luo et al., 2024) comprises 38,803 ICU stays.

We further exclude 65 patients whose 34 time series variables contain no non-zero finite observation within their reported sequence length, yielding 38,738 samples. All 65 carry a positive sepsis label, so removing them lowers the positive rate from 4.19% to 4.03%. Such records provide no measurement signal for any time series classifier and only inflate the apparent positive rate. Because we deviate from the sample set used by prior work, we generate five new splits for P19.

MIMIC-III. MIMIC-III (Johnson et al., 2016) is a publicly released critical care database distributed via PhysioNet.[8] We adopt a benchmark preprocessing pipeline (Harutyunyan et al., 2019) and target binary in-hospital mortality prediction over the first 48 hours after ICU admission. The resulting dataset comprises 21,107 patient stays, each described by 16 time series variables and one static feature (Height).

MIMIC-IV. MIMIC-IV (Johnson et al., 2023) is a critical care database of sparse, irregularly sampled physiological records collected at the Beth Israel Deaconess Medical Center between 2008 and 2019.[9] We follow the preprocessing of GSNF (Gao et al., 2026), extracting 96 features (patient intake and output, laboratory results, and medication prescriptions) from the first 48 hours of recorded measurements, with binary in-hospital mortality as the target. A total of 26,070 hospital admissions are retained. Unlike P12, P19, and MIMIC-III, this cohort carries no static features, as the preprocessing produces none.

eICU. eICU (Pollard et al., 2018) contains records from patients admitted to intensive care units across 208 hospitals in the United States between 2014 and 2015.[10] We follow GSNF (Gao et al., 2026), extracting 14 features from the first 48 hours after ICU admission for the same in-hospital mortality task. A total of 12,312 ICU stays are retained, each belonging to a distinct patient, and, as for MIMIC-IV, no static features are used.

[7] https://physionet.org/content/challenge-2019/1.0.0/
[8] https://physionet.org/content/mimiciii/1.4/
[9] https://physionet.org/content/mimiciv/1.0/
[10] https://physionet.org/content/eicu-crd/2.0/

C.2 METRICS

AUROC and AUPRC follow their standard definitions. The Brier score is the mean squared error between the predicted probability and the binary outcome. ECE (Guo et al., 2017) measures the average gap between predicted confidence and empirical accuracy across $M = 10$ equal-width bins on $[0, 1]$. For both the Brier score and ECE, lower values indicate better calibration.

C.3 BASELINES

We employ three categories of baselines: zero-shot LLMs, ISMTS baselines, and LLM-based prediction methods.

For the zero-shot LLMs, we use GPT-5.1 (OpenAI, 2025) and gpt-oss-120b (Agarwal et al., 2025), both with `reasoning_effort=medium`, and Qwen3-4B (Yang et al., 2025) with `enable_thinking=True`. These include one proprietary model and two open-weight models of different sizes. Based on the preliminary study in Appendix B.3, we use the verbalized probability produced by the model as the predicted risk. The prompts used for zero-shot evaluation are provided in Appendix H.

For all ISMTS baselines, we use Adam as the default optimizer and follow each baseline's released code for model-specific training settings, including the optimizer where the released code uses a different one. To mitigate class imbalance, we follow the training protocol of Zhang et al. (2022); Li et al. (2023); Luo et al. (2024) and upsample the minority class threefold while constraining each mini-batch to a 1:1 class ratio. Because AUPRC is the primary metric for irregular time series classification under class imbalance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), we use validation AUPRC as the model selection criterion. Unless otherwise specified, each model is trained for up to 50 epochs with an early stopping patience of 5. We evaluate each hyperparameter configuration over 5 runs and select the configuration with the highest validation AUPRC averaged over these runs. Method-specific hyperparameter configurations and search spaces are described below.

GRU-D (Che et al., 2018). We reproduce the original authors' Keras implementation[11] in PyTorch with the original MLP classification head. Following the released code, we fix the input and recurrent dropout to 0.3. We tune the hidden dimension over $\{32, 64, 128, 256, 512, 1024\}$, the learning rate over {5e-4, 1e-3, 5e-3, 1e-2}, and the batch size over $\{32, 64, 128, 256, 512\}$.

mTAND (Shukla & Marlin, 2021). We follow the official implementation[12] and adopt the mTAND-Full classification variant. Following the released code, we fix the number of reference time points to 128, the time-embedding dimension to 128, the number of attention heads to 1, and the generator hidden size to 50. We tune the recurrent encoder hidden size over $\{32, 64, 128\}$, the latent dimension over $\{32, 64, 128\}$, the classification-loss weight over $\{1, 5, 10, 50, 100\}$, the learning rate over {5e-4, 1e-3, 5e-3, 1e-2}, and the batch size over $\{32, 64, 128\}$.

SeFT (Horn et al., 2020). We reproduce the authors' TensorFlow/Keras implementation[13] in PyTorch. Following the values dominant across datasets in the original paper's hyperparameter table, we fix the observation encoder to a 4-layer MLP, the attention aggregator to 2 layers with 4 heads, and the output readout to a 2-layer MLP of width 512. We tune the per-observation embedding dimension over $\{256, 512\}$, the observation encoder MLP width over $\{64, 128\}$, the shared dropout rate over $\{0.0, 0.2, 0.4\}$, the time-encoding dimension over $\{8, 16\}$, the maximum time scale over $\{100, 1000\}$, the learning rate over {1e-4, 5e-4, 1e-3, 5e-3}, and the batch size over $\{32, 64, 128\}$.

Raindrop (Zhang et al., 2022). We follow the official implementation.[14] Following the original paper, we fix the batch size to 128. We tune the per-sensor observation embedding dimension over $\{2, 4, 8\}$, the number of Raindrop layers over $\{1, 2, 4\}$, the number of attention heads over $\{1, 2, 4\}$, dropout over $\{0.0, 0.2, 0.4\}$, and the learning rate over {1e-4, 5e-4, 1e-3, 5e-3}.

[11] https://github.com/PeterChe1990/GRU-D
[12] https://github.com/reml-lab/mTAN
[13] https://github.com/BorgwardtLab/Set_Functions_for_Time_Series
[14] https://github.com/mims-harvard/Raindrop

STraTS (Tipirneni & Reddy, 2022). We follow the official PyTorch implementation.[15] Following the released code and paper, we use AdamW with the default weight decay of 0.01 and fix the learning rate to 5e-4, the batch size to 32, and the per-patient observation cap to 880. We tune the hidden dimension over $\{32, 64, 128\}$, the number of Transformer layers over $\{1, 2, 4\}$, the number of attention heads over $\{8, 16, 32\}$, and the shared dropout rate (used for both feature and attention dropout) over $\{0.0, 0.2, 0.4\}$.

ViTST (Li et al., 2023). We follow the official implementation[16] and adopt the Vision-Text variant, which combines a Swin Transformer image encoder with a RoBERTa text encoder; both encoders are fine-tuned end-to-end. Following the original paper, we render the time series inputs as line-graph images at a resolution of 384×384. We use dataset-specific grid layouts according to the number of input variables: 6×6 for P12 and P19, 4×4 for MIMIC-III and eICU, and 12×12 for MIMIC-IV. We set the maximum number of training epochs to 4 for P12, MIMIC-III, MIMIC-IV, and eICU, and to 2 for P19, with an effective batch size of 48. We tune the learning rate over $\{$1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4$\}$ and select both the learning rate and the checkpoint based on validation AUPRC.

KEDGN (Luo et al., 2024). We follow the official implementation.[17] Following the released code, we fix the batch size to 512, 512, and 256 and the learning rate to 1e-3, 5e-3, and 5e-3 for P12, P19, and MIMIC-III, respectively, and use the released ChatGPT-extracted textual descriptions. For MIMIC-IV and eICU, we fix the batch size to 256, tune the learning rate over $\{$1e-3, 5e-3$\}$, and generate the textual descriptions with GPT-5.6 Sol using the prompt provided in the original paper. We use GPT-5.6 Sol in place of ChatGPT because the ChatGPT version used in the original paper is no longer available. The model is used only to generate these variable descriptions, and its input consists of the prompt and the variable names. At no point are raw measurement values or any other patient-level data from MIMIC-IV or eICU sent to the model. On all datasets, the textual descriptions are encoded by BERT. We tune the dimension of the query vectors over $\{5, 7, 9\}$, the dimension of the variable node embeddings over $\{7, 9, 11\}$, the proportion of the density score over $\{1.0, 2.0, 3.0\}$, and the dimension shared by the variable hidden states and the structured encoding representations over $\{8, 12, 16\}$.

Hi-Patch (Luo et al., 2025). We follow the official implementation.[18] Following the released code, we train each model for up to 20 epochs and fix the time-decay coefficient to 1, the learning rate to 1e-3, and the batch size to 8, 64, and 16 for P12, P19, and MIMIC-III, respectively. For MIMIC-IV and eICU, we fix the batch size to 16. We tune the hidden dimension over $\{16, 32, 64, 128\}$, the number of attention heads over $\{1, 2, 4, 8\}$, the number of graph-attention layers over $\{1, 2\}$, and the patch size over $\{T/2, T/4, T/8, T/16, T/32\}$, where the history window T is 60 hours for P19 and 48 hours for the other datasets.

GSNF (Gao et al., 2026). We follow the official implementation.[19] Since the original paper reports hyperparameters for every dataset it evaluates, we adopt them without tuning: the optimizer is Adam with a weight decay of 1e-4, the batch size is 50, the learning rate is 1e-3 decayed by a factor of 0.5 every 20 epochs, the number of GSNF layers is 2, the number of hidden layers is 3, the hidden dimension is 128, the cross-entropy weight is 1000, and the weights of the interaction-aware trajectory generation (ITG) and reverse-time trajectory generation (RTG) losses are both 0.1; the released code sets the latent dimension to the number of variables. The separation margin is dataset-specific: 1e-5 for P12, P19, and eICU, and 1e-7 for MIMIC-IV. MIMIC-III is not evaluated in the original paper, so on this dataset we tune the separation margin over its candidate set $\{$1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8$\}$ and leave all other hyperparameters unchanged. As in the original paper, predictions are taken from the model's own classification head.

[15] https://github.com/sindhura97/STraTS
[16] https://github.com/Leezekun/ViTST
[17] https://github.com/easonLuo2001/KEDGN
[18] https://github.com/easonLuo2001/Hi-Patch
[19] https://github.com/mzgaooo/GSNF

GARLIC (Li et al., 2026). We follow the official implementation.[20] Following the released code, we train each model for up to 100 epochs with an early stopping patience of 10. We switch between the reconstruction and prediction stages whenever the validation loss has not improved for more than 3 epochs, reloading the checkpoint with the lowest validation loss at each switch. Since AUPRC is our primary evaluation metric, we select the final model based on the best validation AUPRC. We fix the reconstruction learning rate to 1e-3, the graph-regularization weight to 0.01, and both weight-decay coefficients to 1e-4. We tune the lag window over $\{3, 5, 7, 9\}$, the batch size over $\{64, 128, 256\}$, the hidden dimension over $\{128, 512\}$, the feature dimension over $\{16, 32\}$, the classification learning rate over $\{$5e-5, 1e-4$\}$, and the classification-loss weight over $\{0.1, 10\}$ for all datasets.

For the LLM-based prediction methods, Record2Vec and TimeCAP, we fix the LLM to Record2Vec's default summarizer, medgemma-27b-text-it (Sellergren et al., 2025); both methods also convert each record into text with the missingness-aware serialization of the Record2Vec code. Training follows the ISMTS protocol above, except that TimeCAP keeps the loss defined in its released code and uses neither the minority-class upsampling nor the 1:1 mini-batch ratio.

Record2Vec (Ji et al., 2026). We follow the official implementation.[21] The LLM summarizes each serialized record once, and the summaries are embedded with Qwen3-Embedding-8B (Zhang et al., 2025b). Following the released code, we set the batch size to 512, the maximum number of epochs to 200, and the early stopping patience to 8. We tune the classification head over $\{$MLP, LSTM, PatchTSMixer$\}$ and the learning rate over $\{$1e-5, 5e-5, 1e-4, 5e-4$\}$.

TimeCAP (Lee et al., 2025). We follow the official implementation.[22] The LLM contextualizes each serialized record once, and the resulting text is encoded by BERT (Devlin et al., 2019)[23], which is fine-tuned end-to-end. Following the released code, we fix the model width to 64 and the feed-forward width to 128 and train for up to 100 epochs with an early stopping patience of 10. We tune the number of attention layers over $\{1, 2\}$, the number of attention heads over $\{4, 8, 16\}$, the dropout rate over $\{0.0, 0.1, 0.2\}$, and the learning rate over $\{$1e-4, 5e-4, 1e-3$\}$. The remaining settings are adapted to the scale and irregular sampling of our datasets: we fix the patch length to 8, the stride to 4, and the batch size to 64; replace instance normalization with per-variable standardization using training-fold statistics; and fill missing values by last observation carried forward. Retrieval uses 5 neighbors without class stratification, and the weight that combines the predictions of the multimodal encoder and the LLM is selected on the validation set of each dataset.

C.4 Implementation Details

Data collection. We collect outcome-specific rationales using GPT-5.1 for P12 and P19, and Kimi K2 Thinking (Team et al., 2025) for MIMIC-III, MIMIC-IV, and eICU. We first collect rationales *separately* for each candidate outcome. For each candidate, the LLM is prompted to assume that outcome and to identify only the patient features that support it. To keep each rationale strictly outcome-specific rather than contrastive, we prohibit the LLM from referencing any alternative outcome or fabricating unobserved evidence; when no supporting evidence appears, the rationale is left blank. These constraints are intended to keep the rationales concise and grounded in observed features, without hallucinated or contrastive content.

To mitigate class imbalance, we follow prior work (Zhang et al., 2022; Li et al., 2023; Luo et al., 2024) and oversample the minority class during data collection. For each minority class sample, we collect three rationales per candidate outcome on P12, MIMIC-III, MIMIC-IV, and eICU, and six on P19, whose positive rate is lower. Majority class samples receive one rationale per candidate outcome. The full set of prompts is provided in Appendix H.

[20] https://github.com/SCAI-Lab/GARLIC
[21] https://github.com/Jerryji007/Record2Vec-ICLR2026
[22] https://github.com/geon0325/TimeCAP
[23] https://huggingface.co/google-bert/bert-base-cased

Table 9: SFT experimental setup details.

Parameter	Value
batch_size	64 (P12) 128 (Others)
max_length	12,288 (P12/MIMIC-III) 11,264 (P19) 16,384 (MIMIC-IV/eICU)
num_train_epochs	3
optimizer	AdamW
learning rate	2×10^{-5}
lr_scheduler_type	cosine
warmup_ratio	0.05
weight_decay	0.01
max_grad_norm	1.0
completion_only_loss	true

Table 10: RL experimental setup details.

Parameter	Value
max_prompt_length	10,240 (P12/MIMIC-III) 8,192 (P19) 16,384 (MIMIC-IV/eICU)
max_response_length	2,048
train_batch_size	256
ppo_mini_batch_size	32
total_training_steps	150
optimizer	AdamW
learning rate	1×10^{-6}
lr_warmup_steps	10
lr_scheduler_type	constant
rollout.n	8
rollout.temperature	1.0
clip_ratio_low	0.20
clip_ratio_high	0.28
kl_loss_coef	0.001
kl_loss_type	mse (k_2)

SFT. Following prior work (Zhang et al., 2022; Li et al., 2023; Luo et al., 2024), we employ a custom sampler that maintains a 1:1 minority-to-majority ratio within each batch. All SFT experiments are conducted on a single H200 GPU. The full set of hyperparameters is listed in Table 9.

RL. For RL initialization, we select the best SFT checkpoint up to two epochs rather than the best checkpoint over all three SFT epochs, as this empirically yields better RL performance and aligns with recent observations that high SFT scores may not predict better post-RL performance (Kang et al., 2026). Following Dr. GRPO (Liu et al., 2025c) and DAPO (Yu et al., 2025), we compute advantages without standard-deviation normalization and use a token-level policy gradient loss. For KL regularization, we adopt the k_2-as-loss formulation (Liu et al., 2025b).

To address class imbalance, the advantage is scaled by a factor of 2 for minority class samples and by $\frac{2}{N}$ for majority class samples, where $N : 1$ denotes the ratio of the majority to minority class size in the full training set. In particular, in the case of the P19 data, the minority class samples are oversampled by a factor of three due to severe class imbalance. We also employ a custom stratified sampler to ensure that each rollout batch, rather than the PPO mini-batch, matches the class ratio of the full training set.

All RL experiments are implemented on top of `verl` (Sheng et al., 2025) and performed on 2× B200 GPUs. The full set of training hyperparameters is in Table 10. We use $m = 1.0$ as the margin in the reward and $\lambda = 0.25$ as the weight of the CE loss in Equation 3. Note that the batch size is set to 256, which is larger than typical configurations, in order to ensure that our batch-level reward operates stably.

Inference. We use a fixed sampling temperature of 0.7 during inference. Because the model is trained on both rationale orderings in Equation 2 to improve generalization, we likewise perform inference under each. For each pass, we apply a softmax to the logits of y^+ and y^- at the $\hat{y}$ position and take the resulting probability of y^+ as the risk score. The patient's final risk score is defined as the mean of these two probabilities. A comparison between this bidirectional averaging and using each ordering individually is provided in Appendix G.3.

D Full Results on Discrimination

Table 11 reports the results of all methods, including the ISMTS baselines omitted from Table 3. It also shows how the natural-language-explainable methods perform relative to the ISMTS baselines, which cannot explain their predictions in natural language.

Table 11: Discrimination performance on clinical risk prediction benchmarks. The best and second-best results among all methods, including the ISMTS baselines, are indicated in **bold** and underline, respectively.

Method	P12		P19		MIMIC-III		MIMIC-IV		eICU	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Specialized ISMTS Models (non-explainable in natural language)										
GRU-D	87.1±0.9	56.3±2.3	**89.7±0.9**	**56.6±2.9**	84.7±0.2	49.0±0.6	87.5±0.5	59.2±1.4	84.4±0.7	56.4±2.2
mTAND	86.8±0.7	56.6±2.4	84.4±1.6	39.4±4.5	84.6±0.3	49.1±0.9	87.0±0.2	58.0±0.8	84.7±0.9	56.9±1.4
SeFT	86.6±1.1	55.7±4.4	88.1±1.2	51.1±2.6	84.7±0.4	48.0±1.2	87.2±0.4	56.8±2.2	85.0±0.5	56.0±0.6
Raindrop	82.6±1.5	45.4±1.2	85.5±1.1	52.2±3.0	81.0±0.6	37.2±1.4	83.3±0.4	47.4±1.1	80.9±0.8	50.8±1.8
STraTS	87.3±1.5	57.4±3.9	88.7±1.4	50.4±2.3	85.3±0.5	48.3±0.6	**88.6±0.4**	61.3±1.1	**85.3±0.7**	57.4±2.6
ViTST	85.1±0.7	50.4±2.7	86.8±1.0	45.4±3.1	81.3±0.7	38.9±0.8	86.4±0.9	57.2±2.6	84.3±1.0	54.4±2.8
KEDGN	**87.8±1.1**	58.3±3.2	88.2±0.5	53.5±3.0	84.7±0.4	48.5±0.9	87.9±0.5	60.1±1.2	**85.3±0.6**	56.4±1.4
Hi-Patch	87.3±0.8	57.0±3.7	88.7±1.6	52.1±4.3	84.6±0.6	46.2±0.3	87.7±0.3	57.7±0.6	85.0±0.8	56.5±2.5
GSNF	84.7±1.3	50.7±2.4	86.2±2.0	46.2±3.2	83.6±0.4	40.5±0.7	86.0±0.5	54.6±1.6	83.4±0.7	54.4±1.1
GARLIC	86.6±0.8	55.3±2.2	88.8±0.6	52.4±1.8	85.5±0.3	50.5±1.5	87.2±0.5	57.9±1.8	85.2±1.0	58.2±1.5
Zero-shot LLMs										
GPT-5.1	83.9±1.0	49.3±0.5	72.2±2.9	9.3±0.9	–	–	–	–	–	–
gpt-oss-120b	81.2±0.7	43.4±3.0	64.2±2.1	6.7±0.6	76.6±0.6	31.0±1.0	70.8±0.5	27.9±1.0	79.5±1.2	47.4±1.0
Qwen3-4B	69.7±1.9	26.7±0.8	56.6±1.7	5.0±0.6	69.3±0.5	20.2±0.2	65.6±0.7	20.9±0.8	71.2±1.4	31.4±0.9
LLM-based Prediction Methods										
Record2Vec	76.9±1.3	38.2±2.9	79.5±1.6	33.4±3.8	73.6±0.1	31.9±0.1	79.5±1.2	43.6±1.9	75.9±0.5	42.5±1.1
TimeCAP	85.2±1.0	52.8±3.9	85.4±1.3	33.7±2.3	84.2±0.1	46.9±0.2	87.3±0.3	57.5±0.5	83.8±1.0	54.6±3.2
TRIAGE$_{\text{SFT}}$	86.9±1.0	56.4±1.9	88.9±1.0	52.2±3.2	86.4±0.3	51.4±0.9	87.5±0.2	59.7±0.8	84.2±1.3	56.7±0.9
TRIAGE	87.3±1.2	**59.0±3.4**	89.3±1.0	53.8±1.4	**86.7±0.2**	**54.1±0.5**	87.8±0.5	**61.6±2.1**	85.2±0.5	**58.7±1.4**

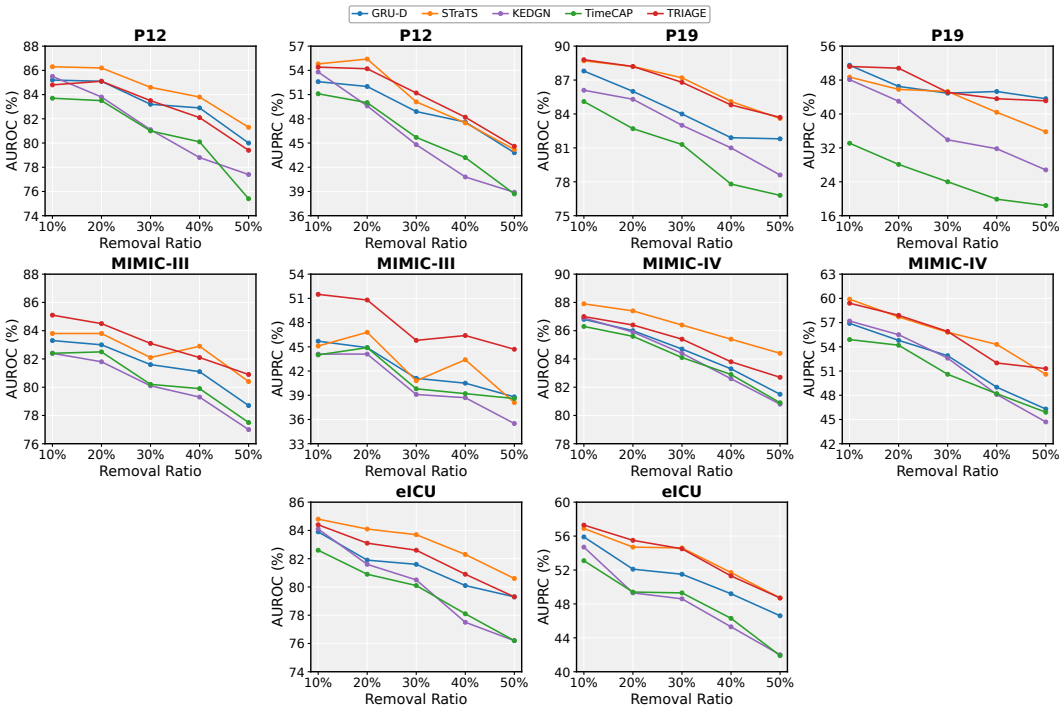

Figure 6: Performance comparison under the leave-random-variables-out setting on five ISMTS benchmarks.

E Robustness under Limited Information

E.1 Leave-Random-Variables-Out

In practice, clinical risk predictors rarely observe every variable at inference time due to sensor failures, protocol differences, or missing records. Prior work has therefore evaluated ISMTS models under *variable-level* missingness, where a subset of variables is entirely removed at inference (Zhang et al., 2022; Li et al., 2023; Luo et al., 2024). We randomly remove 10%, 20%, 30%, 40%, and 50% of the variables and evaluate the discrimination performance to assess the robustness of TRIAGE. We compare TRIAGE against the three top-performing ISMTS models, GRU-D, KEDGN, and STraTS, and the LLM-based method TimeCAP.

Across the five benchmarks, TRIAGE attains the best AUPRC on 15 out of 25 settings, while outperforming TimeCAP on all 50 settings (Figure 6). Crucially, TRIAGE is the least sensitive to variable removal: averaged over the ten dataset–metric pairs, its performance drops by 6.6 points (10.4% relative) as the removal ratio increases from 10% to 50%, compared with 6.9 points for GRU-D (11.3%), 6.9 points for STraTS (11.6%), 10.5 points for KEDGN (17.8%), and 8.6 points for TimeCAP (15.8%). These results indicate that TRIAGE does not depend on a specific subset of variables and retains its predictive signal when the observed variable set is substantially reduced.

E.2 Low-Resource Training

Clinical applications frequently suffer from limited labeled data, particularly for rare conditions and newly emerging diseases. These regimes favor models that can draw on pretrained medical and reasoning priors. Unlike ISMTS baselines trained from scratch, TRIAGE builds on a pretrained LLM and inherits such priors directly, without requiring large amounts of task-specific supervision.

Setup. To probe this advantage under data scarcity, we fine-tune TRIAGE with rsLoRA (Kalajdzievski, 2023), which preserves pretrained representations more reliably than full fine-tuning in low-data regimes. We use rank $r = 64$, scaling factor $\alpha = 64$, and dropout 0.05, applying adapters to all linear layers. We omit the RL stage to isolate the data efficiency of supervised reasoning alone. Only the P12 training set is subsampled, to 1% and 5% of its original size; the validation and test sets are kept intact for consistent evaluation.

Baseline tuning. For a fair comparison, we start from the hyperparameters of the full-data setting and additionally sweep smaller batch sizes and a broader range of learning rates. This ensures that observed differences reflect data efficiency rather than under-tuned baselines.

Table 12: Results on P12 under low-resource training. Only the training set is subsampled. The best and second-best results are indicated in **bold** and underline, respectively.

	P12-1%		P12-5%	
Method	AUROC	AUPRC	AUROC	AUPRC
GRU-D	74.5±4.2	35.3±5.3	**82.3**±1.1	<u>48.0</u>±2.0
STraTS	71.8±5.2	29.8±7.3	80.8±2.0	42.5±3.5
KEDGN	69.4±5.7	28.1±8.0	80.8±1.5	45.5±2.3
TimeCAP	**76.4**±1.6	<u>37.9</u>±3.2	80.3±1.7	45.5±2.5
TRIAGE$_{\text{SFT}}$	<u>76.2</u>±4.8	**40.0**±5.6	**82.3**±1.1	**49.0**±3.6

Results. We compare TRIAGE against GRU-D, STraTS, and KEDGN, the three ISMTS models with the best average performance in our main experiments, and the LLM-based method TimeCAP (Table 12). LLM-based methods are less sensitive to the reduced label budget: from P12-5% to P12-1%, TRIAGE and TimeCAP lose 9.0 and 7.6 AUPRC, while GRU-D, STraTS, and KEDGN lose 12.7, 12.7, and 17.4, respectively. This is consistent with the hypothesis that pretrained knowledge contributes most when supervision is scarce. TimeCAP, however, plateaus at a lower operating point, and TRIAGE surpasses it in AUPRC at both budgets, by 5.5% and 7.7% relative at P12-1% and P12-5%, respectively. Against GRU-D, the strongest of these ISMTS baselines, the margin is larger when labels are scarcer: TRIAGE improves AUPRC by 13.3% relative at P12-1% and by 2.1% relative at P12-5%.

Table 13: Expert evaluation of generated reasoning. Two clinicians rate all rationales on a 1–5 scale. $^{\dagger}/^{\ddagger}$: differs from TRIAGE at $p < 0.05$ / $p < 0.01$ (two-sided Wilcoxon signed-rank over cases).

Method	Consistency	Correctness	Helpfulness	Overall
STraTS + IG (textualized)	3.28‡	3.43‡	2.48‡	3.07‡
GPT-5.1 (zero-shot)	**4.72**	**4.60**‡	**4.73**†	**4.68**‡
TRIAGE (one-sided reasoning)	4.43	4.32	3.93‡	4.23‡
TRIAGE	4.55	4.35	4.55	4.48

F Additional Analysis of TRIAGE Reasoning

F.1 Human Expert Evaluation

Protocol. We ask two clinicians to conduct an expert evaluation of the reasoning generated by TRIAGE. From the P12 test split, we draw 30 ICU stays (20 survivors, 10 non-survivors) and present four rationales for each case, giving 120 reasoning chains in total: (i) TRIAGE; (ii) a *one-sided rationale* described in Section 7.1; (iii) zero-shot GPT-5.1; and (iv) the textualized Integrated Gradients (IG; Sundararajan et al., 2017) attributions over STraTS (Tipirneni & Reddy, 2022) described in Section 7.2. The IG baseline is truncated to the length of the corresponding TRIAGE rationale so that verbosity cannot by itself drive the ratings. Each rationale is rated on consistency, correctness, and helpfulness using the 1–5 rubric (Figure 26). We treat the case as the unit of analysis, averaging the two raters, and test differences with two-sided Wilcoxon signed-rank tests. Table 13 reports the results.

Dialectical reasoning is preferred, specifically for decision support. Against the one-sided reasoning, TRIAGE is indistinguishable on consistency ($+0.12$, $p = 0.19$) and correctness ($+0.03$, $p = 0.66$) but clearly higher on helpfulness ($+0.62$, 95% CI $[+0.35, +0.88]$, $d_z = 0.80$, $p < 0.001$). This isolates the dialectical format's contribution to perceived utility. Displaying the competing outcome states which findings argued against the prediction and why they were outweighed, which is the information a clinician needs to audit the model's judgment, and which the one-sided rationale removes. Together with the discrimination gains of Table 5, this indicates that reasoning over competing outcomes improves the interpretability and real-world helpfulness it was introduced to provide.

Attribution-based and frontier-model explanations. Textualized IG scores lowest on all three criteria, and its deficit is larger on helpfulness (2.48) than on consistency (3.28) or correctness (3.43). The attributions report which measurements moved the prediction, but they do not state the physiology that connects them, and the clinician is left to reconstruct the inferential step. This is consistent with prior reports that saliency-style explanations transfer poorly to clinical use (Hou & Wang, 2025). Zero-shot GPT-5.1 receives the highest ratings overall (4.68 against 4.48), by a small margin that is significant on two of the three criteria (consistency -0.17, $p = 0.18$; correctness -0.25, $p < 0.01$; helpfulness -0.18, $p < 0.05$). We attribute the remaining gap to capacity, since TRIAGE uses a 4B backbone. Crucially, these ratings do not carry over to risk estimation: the same GPT-5.1 configuration reaches 49.3 AUPRC on P12 against 59.0 for TRIAGE (Table 3). A rationale that accompanies a weaker prediction is of limited clinical use, so the comparison that matters is whether reasoning quality can be obtained together with calibrated discrimination at a deployable model scale.

Inter-rater agreement. Over the 120 rationales, ICC(A,1) between the two clinicians is 0.61 for helpfulness, 0.35 for consistency, and 0.11 for correctness. The low value for correctness likely reflects, in part, a strong ceiling effect, since one rater gave the maximum score to 87.5% of rationales (Gwet's AC2 $= 0.59$). The raters also differ in strictness (mean 3.72 vs. 4.51), but z-scoring each rater leaves the ordering of methods unchanged on every criterion. Both raters assign higher helpfulness scores to TRIAGE than to the one-sided rationale ($+0.93$, $p < 0.001$; $+0.30$, $p = 0.14$), whereas the advantage of GPT-5.1 comes from one rater alone.

Table 14: LLM-as-a-judge evaluation of clinical reasoning quality on 1,000 P12 samples. Baseline denotes STraTS + IG (textualized). *LLM int.*: GPT-5.1 interpretation of IG attributions. Maximum scores for I/D/E/A are 4/2/2/2, respectively.

Criterion	Baseline	Baseline+LLM int.	TRIAGE
Interpretive summary (I)	1.384	2.619	**3.425**
Differential diagnosis (D)	0.751	**1.254**	1.213
Lead diagnosis exp. (E)	0.657	1.155	**1.182**
Alternative diagnosis exp. (A)	1.171	1.652	**1.887**
Total	3.964	6.680	**7.708**

Limitations. Because expert annotation by practicing clinicians is costly, our human evaluation is limited to 30 patient cases and two raters, which constrains the generality of the conclusions. Scaling the study to more raters and a larger, more diverse set of cases is left for future work.

F.2 Quantitative Analysis

Expert evaluation does not scale. The study in Appendix F.1 required two clinicians to read 120 rationales, which limits it to a few dozen cases. To obtain an approximate measure of reasoning quality at the scale of the test set, we adopt an LLM-as-a-judge evaluation, and then check its verdict against the clinicians on the cases for which both are available.

Protocol. We evaluate 1,000 randomly sampled P12 cases. As a baseline, we apply IG to STraTS, a strong ISMTS model (Table 11), and render the attributions in two ways: the textualized form shown to the clinicians in Appendix F.1, and an additional GPT-5.1 interpretation step that translates pointwise attributions into temporal trends, which IG alone cannot express and which strengthens the baseline. We first give LLM judges the rubric used by the clinicians in Appendix F.1, but the resulting scores discriminate poorly between methods. We therefore adopt the IDEA assessment tool (Baker et al., 2015; Schaye et al., 2022): each LLM judge evaluates reasoning traces along four dimensions, **I**nterpretive summary, **D**ifferential diagnosis, **E**xplanation of lead diagnosis, and **A**lternative diagnosis explained, with maximum subscores of 4/2/2/2, respectively, summing to a 0–10 total score. We employ three judge models: GPT-5.1 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), and Gemini 3 Flash (Google, 2025). We query each judge three times per sample and report the mean of the nine resulting scores.

Results. TRIAGE achieves a higher total IDEA score than the IG baseline even when the attributions are additionally interpreted by GPT-5.1 (7.708 vs. 6.680; Table 14) and is preferred on 83.5% of cases. The largest gain appears on the interpretive summary criterion (+0.806), indicating that the model captures the patient's baseline risk and temporal trajectory more effectively. The alternative diagnosis dimension also improves (+0.234), in line with our design choice to reason over alternative outcomes, while the scores on the remaining criteria are comparable. Without the additional LLM interpretation step, the textualized IG attributions score 3.964 and TRIAGE is preferred on all cases, consistent with the margin observed in the expert study.

Agreement with the expert evaluation. The human evaluation protocol and LLM-as-a-judge evaluation protocol use different rubrics and different score ranges, so we compare the preferences their scores induce rather than the scores themselves. For each case, we derive a preference for TRIAGE, a tie, or IG from the difference in judge totals, and separately from the difference in each human criterion, and measure the agreement between them. The two agree in every case on consistency and on helpfulness ($30/30$, 95% CI $[88.6\%, 100\%]$), and in 28 of 30 on correctness. Despite this agreement, we recognize that the automatic protocol cannot substitute for evaluation by clinical experts, and we therefore report it only as a scalable proxy for reasoning quality.

F.3 Hallucination Analysis

We employ GPT-5.1 as a judge to detect severe hallucinations in TRIAGE reasoning traces, such as references to features or values that were not present in the patient record (see Figure 32 for the full prompt).

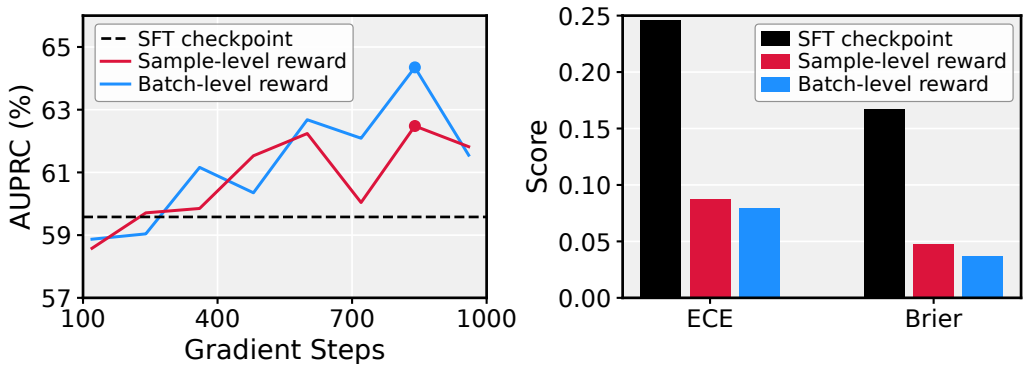

Figure 7: Discrimination (left) and calibration performance (right) under different reward designs on P12.

Out of 1,000 samples from the P12 dataset, only 1.3% exhibit severe hallucinations. Representative examples are in Figure 12. This indicates that during dialectical reasoning, the model rarely fabricates evidence for the opposing side when the signal from one side is already clear-cut. We attribute this favorable outcome to the safeguards built into our data construction process, in particular, the explicit instruction to leave a field blank whenever no supporting evidence is available.

G Further Experiments

G.1 Ablation Study on RL Reward Design

We compare our batch-level reward against a conventional sample-level reward, in which each rollout receives $R_{i,j}^{\text{sample}} = \log P(y_i \mid x_i, r_{i,j})$, followed by the same group normalization. This reward provides a direct correctness signal for each patient, akin to the standard outcome-based reward in GRPO, but does not explicitly encourage risk estimates that are comparable across patients.

Relative to the SFT checkpoint, both rewards improve discrimination (AUPRC 59.58 $\rightarrow$ 62.48 and 64.35) and calibration, reducing ECE from 0.246 to below 0.09 and the Brier score from 0.167 to below 0.05 (Figure 7), which confirms that reinforcement learning itself is a major source of calibration gain and further improvement of discrimination. Compared with the sample-level reward, the batch-level reward improves both discrimination with higher AUPRC (64.35 vs. 62.48) and calibration with lower ECE (0.079 vs. 0.087) and Brier score (0.037 vs. 0.047). These results confirm that the inter-sample signal introduced by the batch-level reward provides a complementary learning objective that improves both the discrimination and calibration of predicted risk scores.

G.2 Ablation Study on the Backbone

To verify robustness to backbone choice, we apply TRIAGE to Qwen3-1.7B and 8B (Yang et al., 2025) in addition to our 4B default, and to Llama-3.2-3B (Grattafiori et al., 2024) from a different model family. As shown in Table 15, scaling Qwen3 from 1.7B to 4B yields a clear AUPRC gain (53.8 to 56.4), while further scaling to 8B provides comparable performance (56.0), motivating our choice of 4B as the default backbone. Across model families, Qwen3-4B achieves higher AUPRC than the similarly sized Llama-3.2-3B (56.4 vs. 54.5), while TRIAGE retains a consistent advantage over the corresponding baseline on every backbone, suggesting that our reasoning supervision generalizes beyond a particular model scale or architecture.

G.3 Inference Direction Analysis

Since TRIAGE is trained on both rationale orderings, we examine whether averaging the risk scores from the two directions at inference offers any benefit over using a single ordering. Table 16 reports the results on P12 with Qwen3-4B-Base under the SFT setting. The three strategies perform comparably, indicating that a single-direction pass alone already yields a reliable risk estimate and can

Table 15: Ablation on the backbone, on P12.

Backbone	AUROC	AUPRC
Qwen3-4B (default)	86.9±1.0	**56.4±1.9**
Scale (Qwen3 family)		
Qwen3-1.7B	86.4±1.0	53.8±1.6
Qwen3-8B	86.8±0.8	56.0±2.1
Architecture		
Llama-3.2-3B	**87.0±1.0**	54.5±1.3

Table 16: Ablation on the inference direction, on P12 with Qwen3-4B-Base under the SFT setting. *Positive-first* and *Negative-first* denote inference using only the (y^+, y^-) or (y^-, y^+) ordering, respectively. *Bidirectional* averages the risk scores from both orderings.

Inference strategy	AUROC	AUPRC
Positive-first only	86.8±1.0	56.1±1.0
Negative-first only	86.8±1.0	56.2±2.2
Bidirectional (ours)	**86.9**±1.0	**56.4**±1.9

serve as a lower-cost alternative. Nonetheless, bidirectional averaging consistently provides a small additional gain (*e.g.*, AUPRC of 56.4 vs. 56.1/56.2), and we adopt it as the default inference strategy.

H PROMPTS

We provide a comprehensive list of all the prompts utilized in our study, organized by their role in the pipeline.

Preliminary study. To motivate our method, we first probe zero-shot LLM behavior on ICU mortality prediction. The default direct-answer prompt is shown in Figure 13, its verbalized-probability variant in Figure 14, and the final evaluation prompt for zero-shot LLMs in Figure 15. The LLM-as-a-judge prompt used to detect whether a chain-of-thought ends with an explicit verdict-closure sentence is shown in Figure 16.

Outcome-conditioned rationale generation. To construct supervision for TRIAGE, we elicit outcome-conditioned rationales from a strong LLM, instantiating one prompt per candidate outcome per patient. The prompt template for the four in-hospital mortality datasets (P12, MIMIC-III, MIMIC-IV, and eICU) is shown in Figure 17, and the corresponding template for P19 (early sepsis prediction) is shown in Figure 18.

Main SFT prompts for TRIAGE. TRIAGE is supervised to produce two outcome-conditioned rationale blocks followed by a single final decision. To control for ordering effects, we use two symmetric variants per dataset. For the in-hospital mortality datasets (P12, MIMIC-III, MIMIC-IV, and eICU), the negative-class-first and positive-class-first variants are shown in Figure 19 and Figure 20, respectively. For P19, the corresponding variants are shown in Figure 21 and Figure 22.

Ablation: one-sided reasoning. For the reasoning-structure ablation (Table 5), we use a single-rationale SFT prompt that retains only the rationale aligned with the ground-truth label. The P12 template is shown in Figure 23.

Faithfulness analysis. For the input-removal experiment (Figure 3), GPT-5.1 locates the evidence behind the rationale sentences selected at each deletion ratio (25%, 50%, 75%, and 100%) and returns, for each sentence, the features it relies on together with the value range, time window, or whole series to erase; a deterministic script then removes the matching measurements from the input. This prompt is shown in Figure 24. For the regeneration test at a deletion ratio of 50%, a GPT-5.1 judge compares every sentence of the original rationale with the rationale regenerated from the perturbed input and scores it from 1 (retained) to 5 (absent), with intermediate scores for claims restated with different values, findings kept without their values, and partial retention of the cited evidence. The judge prompt is shown in Figure 25.

Human expert evaluation criteria. The rubric presented to the clinical experts is adapted from the rationale evaluation protocol of Jiang et al. (2025a); the definitions of the criteria and the associated rating scales are shown in Figure 26. To reduce redundancy among evaluation dimensions, we

omit human-likeness and specificity, which capture aspects of rationale quality that partially overlap with helpfulness, and focus on consistency, correctness, and helpfulness as complementary measures of grounding, medical validity, and clinical utility.

Reasoning analysis: XAI interpretation. To compare model-internal explanations with LLM-style rationales, we textualize attribution-based evidence from STraTS and ask an LLM to interpret it in the same two-sided rationale format. The interpretation prompt is shown in Figure 27.

Quantitative analysis: IDEA-based reasoning quality assessment. To assess the clinical quality of generated reasoning traces, we adapt the IDEA rubric into an LLM-as-a-judge protocol that scores Interpretive summary (I), Differential diagnosis (D), Explanation of the lead diagnosis (E), and Alternative diagnosis explained (A). The full evaluation prompt, including task setup, factuality rules, per-domain scoring anchors, overall calibration guidance, and the required JSON output schema, is shown across Figures 28–31.

Hallucination analysis. To quantify hallucination in our reasoning traces, we adopt an LLM-as-a-judge approach that evaluates whether each trace references features or values absent from the patient record. We use GPT-5.1 as a judge, and the exact prompt is in Figure 32.

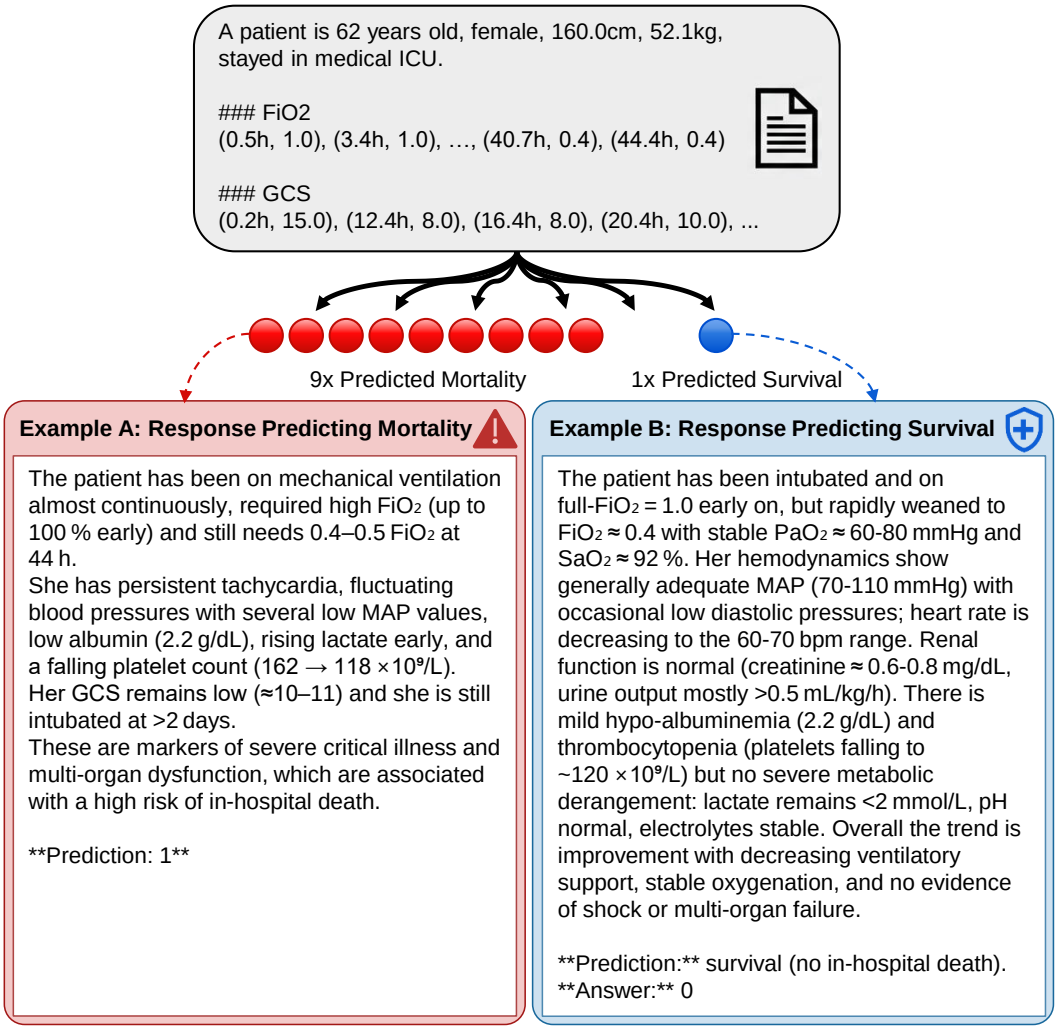

Figure 8: Two reasoning traces sampled from gpt-oss-120b on the same P12 patient. The mean implicit mortality probability over 10 samples for this patient is 0.9. Each trace cites only one-sided evidence and commits to the corresponding outcome.

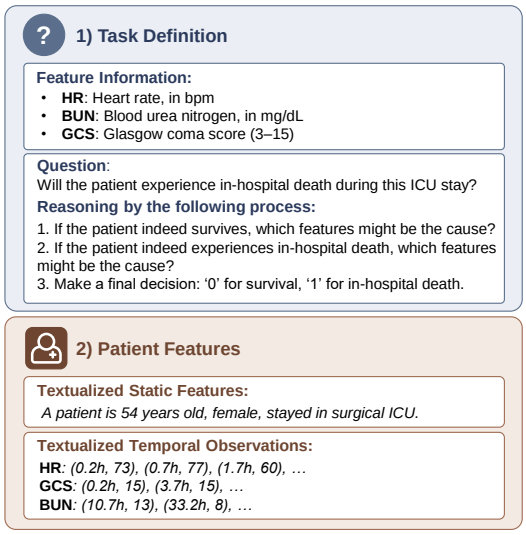

Figure 9: Example prompt format used by TRIAGE. Only some of the features are shown. The input is organized into task definition and textualized patient features.

Patient # 148874		Ground Truth: Survival

A patient is 85 years old, male, stayed in surgical ICU.
For each feature, measurements are listed as (Time, Value) pairs in chronological order, where Time denotes hours since ICU admission.

GCS
(0.2, 15.0), (1.7, 15.0), (5.7, 15.0), (9.7, 15.0), (13.7, 15.0), (17.7, 15.0), (21.7, 15.0), (25.7, 15.0), (29.7, 15.0), (33.7, 15.0), (37.7, 15.0), (41.7, 15.0), (45.7, 15.0)
DiasABP
(4.7, 56.0), (5.0, 58.0), (5.5, 54.0), (5.7, 53.0), (6.7, 52.0), (7.7, 51.0), (8.7, 47.0), (9.7, 49.0), (10.7, 52.0), (11.7, 46.0), (12.7, 51.0), (13.7, 47.0), (14.7, 50.0), (15.7, 48.0), (16.7, 46.0), (17.7, 67.0), (18.7, 59.0), (19.7, 59.0), (20.7, 57.0), (21.7, 58.0), (22.7, 61.0), (23.7, 54.0), (24.7, 54.0), (25.7, 61.0), (26.7, 66.0), (27.7, 57.0), (28.7, 63.0), (29.7, 64.0), (30.7, 66.0), (32.7, 71.0), (33.7, 65.0), (35.2, 69.0), (35.7, 59.0), (37.7, 62.0), (38.7, 62.0), (39.7, 65.0), (40.7, 66.0), (41.5, 65.0), (41.7, 67.0), (42.0, 66.0), (42.2, 65.0), (42.5, 71.0), (42.7, 68.0), (43.7, 55.0), (44.7, 59.0), (45.7, 54.0), (46.7, 52.0), (47.7, 62.0)
HCO3
(5.7, 19.0), (12.6, 21.0), (31.7, 20.0)
Lactate
(5.9, 1.4), (31.9, 1.6)
MechVent
(35.7, 1.0), (38.2, 1.0), (40.7, 1.0), (42.7, 1.0)
Temp
(0.2, 34.9), (0.7, 35.9), (3.7, 35.6), (8.7, 35.8), (13.7, 35.8), (17.7, 36.1), (21.7, 36.2), (25.7, 36.0), (30.7, 35.7), (34.7, 35.8), (37.7, 35.6), (42.7, 35.0), (45.7, 35.3)
Urine
(0.2, 100.0), (1.7, 20.0), (2.7, 15.0), (3.7, 15.0), (4.7, 5.0), (5.7, 10.0), (6.7, 20.0), (7.7, 5.0), (8.7, 15.0), (9.7, 17.0), (10.7, 30.0), (11.7, 10.0), (12.7, 10.0), (13.7, 30.0), (14.7, 25.0), (15.7, 15.0), (16.7, 30.0), (17.7, 30.0), (18.7, 20.0), (19.7, 15.0), (20.7, 20.0), (21.7, 25.0), (22.7, 20.0), (23.7, 30.0), (24.7, 15.0), (25.7, 33.0), (26.7, 30.0), (27.7, 120.0), (28.7, 120.0), (29.7, 160.0), (30.7, 80.0), (32.7, 300.0), (35.5, 100.0), (36.7, 160.0), (37.7, 400.0), (38.7, 100.0), (39.7, 220.0), (40.7, 400.0), (41.7, 100.0), (42.7, 120.0), (43.7, 160.0), (44.7, 320.0), (45.7, 300.0), (46.7, 160.0), (47.7, 100.0)

STraTS + IG		Prediction: Mortality (p_1 = 0.746)	Ground Truth: Survival

Supporting Evidence of in-hospital mortality
Age, 85.0
ICUType, surgical ICU
Temp, (0.2, 34.9), (42.7, 35.0), (45.7, 35.3), (37.7, 35.6), (3.7, 35.6), (34.7, 35.8), (30.7, 35.7), (0.7, 35.9), (8.7, 35.8), (13.7, 35.8)
GCS, (0.2, 15.0), (1.7, 15.0), (45.7, 15.0), (41.7, 15.0), (5.7, 15.0)

Counter-evidence
Urine, (40.7, 400.0), (37.7, 400.0), (44.7, 320.0), (45.7, 300.0)
DiasABP, (11.7, 46.0), (16.7, 46.0), (8.7, 47.0), (13.7, 47.0)
Lactate, (31.82, 1.6)
HCO3, (5.68, 19.0)

STraTS + IG	(LLM Interpretation)	Prediction: Mortality (p_1 = 0.746)	Ground Truth: Survival

Rationale for in-hospital death
- The model viewed the patient's advanced age (85 years) and being in a surgical ICU as baseline factors associated with higher in-hospital mortality risk.
- It relied on repeatedly low body temperatures, mostly between 34.9–35.9 °C from very early (0.2 h) through about 45.7 h, as evidence pointing toward higher risk.
- Repeatedly normal Glasgow Coma Scale scores of 15 were included by the model in the mortality-supporting set.

Rationale for survival
- The model identified several episodes of moderate urine output (300–400 mL at multiple times between ~37.7 and 45.7 h) as favorable, suggesting preserved kidney output.
- Lactate was low at 1.6 mmol/L, which the model considered a positive sign.
- Bicarbonate at 19 mmol/L was also included in the survival-supporting set.

TRIAGE	Prediction: Survival (p_1 = 0.086)	Ground Truth: Survival

Rationale for survival
- Consistently intact neurological status with Glasgow Coma Score 15 throughout the ICU stay, indicating preserved consciousness and cerebral function.
- Lactate levels low (1.4–1.6 mmol/L), indicating absence of significant global tissue hypoperfusion or shock.
- Renal function relatively preserved: creatinine mildly elevated but stable (1.6–1.7 mg/dL) and BUN moderately elevated but stable; urine output improves substantially over time to high hourly volumes, indicating recovery or maintenance of kidney perfusion and function.

Rationale for in-hospital death
- Advanced age (85 years), which is independently associated with higher in-hospital mortality risk during critical illness.
- Persistently low temperature around 35–36 °C (e.g., 34.9–36.2 °C throughout the stay), suggesting hypothermia, which in elderly ICU patients can indicate severe systemic illness, impaired thermoregulation, or sepsis and is associated with worse outcomes.
- Mildly reduced bicarbonate (HCO3 19–21 mmol/L), suggesting a mild metabolic acidosis or reduced buffering capacity, which in the context of critical illness can be associated with increased mortality risk.
- Need for mechanical ventilation later in the ICU course (MechVent=1 at multiple time points), indicating respiratory failure or inability to maintain adequate ventilation/oxygenation without support, a major marker of clinical deterioration and higher mortality.

Figure 10: Case study of a surviving patient. For brevity, only a subset of patient characteristics and some of the clinical reasoning parts are excerpted. Text highlighted in blue and red indicates medically favorable and unfavorable signs, respectively.

Patient # 138409		Ground Truth: Mortality

A patient is 88 years old, male, 52.0 kg, stayed in medical ICU.
For each feature, measurements are listed as (Time, Value) pairs in chronological order, where Time denotes hours since ICU admission.

GCS
(2.0, 7.0), (4.5, 3.0), (8.5, 3.0), (13.5, 5.0), (16.5, 6.0), (20.5, 6.0), (23.5, 6.0), (28.5, 6.0), (32.5, 6.0), (40.5, 9.0), (44.5, 9.0)
BUN
(2.6, 99.0), (9.6, 98.0), (32.4, 82.0), (44.6, 80.0)
Creatinine
(2.6, 2.9), (9.6, 2.3), (32.4, 2.1), (44.6, 2.2)
HCO3
(2.6, 21.0), (2.6, 23.0), (9.6, 20.0), (32.4, 15.0), (44.6, 17.0)
K
(2.6, 6.8), (2.6, 10.0), (9.6, 4.3), (32.4, 4.0), (44.6, 3.8)
TroponinT
(9.6, 0.15)
WBC
(2.6, 21.4), (9.6, 17.9), (9.6, 18.9), (32.4, 23.3)
Na
(2.6, 157.0), (2.6, 160.0), (9.6, 163.0), (21.0, 164.0), (28.6, 160.0), (32.4, 159.0), (44.6, 150.0)

STraTS + IG		Prediction: Mortality (p_1 = 0.977)	Ground Truth: Mortality

Supporting Evidence of in-hospital mortality:
Age, 88.0
Na, (32.35, 159.0), (28.55, 160.0), (20.97, 164.0)
BUN, (44.58, 80.0), (32.35, 82.0)
GCS, (44.43, 9.0), (32.43, 6.0), (40.43, 9.0), (28.43, 6.0), (23.43, 6.0), (20.43, 6.0), (16.43, 6.0), (13.43, 5.0), (8.43, 3.0), (1.93, 7.0), (4.43, 3.0)
ICUType, medical ICU
Weight, 52.0

Counter-evidence:
WBC, (32.35, 23.3)
BUN, (2.52, 99.0), (9.52, 98.0)
Na, (2.52, 160.0)
Gender, male

STraTS + IG	(LLM Interpretation)	Prediction: Mortality (p_1 = 0.977)	Ground Truth: Mortality

Rationale for in-hospital death
- The model places strong weight on the patient's advanced age (88 years) and admission to a medical ICU, both of which it associates with higher in-hospital mortality.
- It relies heavily on repeatedly very low Glasgow Coma Scores (GCS 3–9 across many time points), indicating persistently poor neurological responsiveness.
- Multiple very high sodium values (Na 159–164 mmol/L at several times) are treated as important adverse findings.
- Kidney-related measures support the death prediction: elevated BUN values (around 80–82 mg/dL later in the stay).

Rationale for survival
- The model treats being male as slightly favoring survival in its learned patterns.
- A very high white blood cell count (23.3) is used as counter-evidence to death by the model.
- Some extremely high early BUN values (98–99 mg/dL) and an early very high sodium level (Na 160 mmol/L at 2.52 hours) are also treated as counter-evidence.

TRIAGE	Prediction: Mortality (p_1 = 0.844)	Ground Truth: Mortality

Rationale for in-hospital death
- Advanced age (88 years), which is independently associated with high ICU mortality risk.
- Marked acute kidney injury/renal dysfunction: very high BUN (peak ~99 mg/dL, remaining 80–98 mg/dL) and elevated creatinine (2.1–2.9 mg/dL), indicating severe renal impairment.
- Severe hyperkalemia early in the stay (K up to 10.0 mmol/L; later 6.8 mmol/L), a life-threatening electrolyte disturbance predisposing to fatal arrhythmias.
- Hypernatremia (Na 150–164 mmol/L for much of the stay), suggesting significant fluid/electrolyte derangement and osmotic stress, associated with poor outcomes in the elderly.
- Metabolic acidosis/low bicarbonate (HCO3 as low as 15–17 mmol/L) indicating significant systemic derangement and severity of illness.
- Depressed and fluctuating neurologic status: very low GCS values (as low as 3; later only partial improvement to 6 then 9), indicating severe brain dysfunction and poor neurologic prognosis.
- Leukocytosis (WBC 17.9–23.3) suggesting a significant inflammatory or infectious process, which in this context is associated with high severity of illness.
- Elevated Troponin T (0.15 ng/mL), indicating myocardial injury, which in an elderly, critically ill patient increases the risk of cardiac complications and death.

Rationale for survival
- Glasgow Coma Score improves over time from very low values (3–6) to 9 by around 40–44 hours, indicating neurological recovery.
- Renal function markers (BUN and creatinine) show a downward trend over time, consistent with partial improvement in kidney injury.
- Serum potassium rapidly normalizes from markedly abnormal early values to within acceptable limits, indicating effective correction of a life-threatening electrolyte disturbance.
- Sodium levels, though initially markedly elevated, show a gradual decline over time toward more normal values, suggesting progressive management of hypernatremia.

Figure 11: Case study of a non-survivor. For brevity, only a subset of patient characteristics and some of the clinical reasoning parts are excerpted. Text highlighted in blue and red indicates medically favorable and unfavorable signs, respectively.

Patient # 144683		Ground Truth: Survival

A patient is 34 years old, male, 177.8 cm, 83.0 kg, stayed in surgical ICU.
For each feature, measurements are listed as (Time, Value) pairs in chronological order, where Time denotes hours since ICU admission.

(No Troponin information in this patient)

TRIAGE	Prediction: Survival (p_1 = 0.060)	Ground Truth: Survival

Rationale for survival
- **Hemodynamic and cardiac stability:** Heart rate fluctuates but generally remains in a tolerable range, and *troponin values are not elevated (no data suggesting myocardial injury), consistent with maintained cardiovascular stability.*

(omitted for brevity)

Rationale for in-hospital death
(omitted for brevity)

Patient # 142405		Ground Truth: Survival

A patient is 50 years old, male, 182.9 cm, 127.0 kg, stayed in coronary care unit.
For each feature, measurements are listed as (Time, Value) pairs in chronological order, where Time denotes hours since ICU admission.

GCS
(2.3, 15.0), (5.3, 15.0), (8.3, 15.0), (10.3, 15.0), (16.3, 10.0), (20.3, 10.0), (23.3, 10.0), (29.3, 8.0), (32.3, 10.0), (39.3, 8.0), (47.3, 3.0)

(omitted for brevity)

TRIAGE	Prediction: Tie (p_1 = 0.500)	Ground Truth: Survival

Rationale for survival
- Glasgow Coma Score was 15 early in the stay and remained ≥8 later (with intermittent decreases), indicating preserved neurological function and the ability to protect airway at least at times.

(omitted for brevity)

Rationale for in-hospital death
(omitted for brevity)

Figure 12: Case study of patients with hallucinated clinical rationales. For brevity, only relevant parts are excerpted. In the case of the first patient, the feature (Troponin T) not present in the patient's information is mentioned. In the case of the second patient, a value present in the patient's information is misidentified.

Default prompt used in the preliminary study

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Read the patient's ICU features and answer immediately.

Task
Will the patient experience in-hospital death during this ICU stay?

Reply with only one number, immediately.
- 0 for survival
- 1 for in-hospital death
Do not provide any explanation or additional text.

Feature of the patient
{PATIENT_FEATURES}

Figure 13: Default prompt used in the preliminary study for in-hospital mortality prediction.

Prompt for verbalized probability extraction

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Read the patient's ICU features and answer immediately.

Task
Estimate the probability that this patient will experience in-hospital death during this ICU stay.

Reply with only a single number between 0 and 1, immediately.
- 0.0 means certain survival
- 1.0 means certain death
Do not provide any explanation or additional text.

Feature of the patient
{PATIENT_FEATURES}

Figure 14: Prompt used for verbalized probability extraction. Only the `## Task` section differs from Figure 13.

Prompt used for zero-shot LLM evaluation

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Task
Estimate the probability that this patient will experience in-hospital death during this ICU stay.

Analyze the patient's clinical state, then answer.

Reply with only a single number between 0 and 1 at the end.
- 0.0 means certain survival
- 1.0 means certain death

Feature of the patient
{PATIENT_FEATURES}

Figure 15: Final prompt used for zero-shot LLM evaluation of in-hospital mortality.

Prompt used for LLM-as-a-judge verdict-closure detection

You will read a chain-of-thought reasoning produced by another LLM on a binary in-hospital mortality task (0 = survival, 1 = in-hospital death).

Strict criterion:
The VERY LAST SENTENCE of the reasoning must, by itself, explicitly state the final binary answer — either:
(a) the digit "0" or "1" in the context of the answer, OR
(b) the words "survival"/"survive"/"live" or "death"/"die" presented as the final decision.

If the last sentence does anything else — discusses evidence, raises a caveat, gives a meta-instruction (e.g. "Thus output single number"), asks a question, or trails off — output 0.

Reply with exactly one digit:
- 1 = the LAST sentence is itself an explicit confirmatory restatement of the final answer
- 0 = it is not

Reasoning to evaluate:
"""
{reasoning}
"""

Figure 16: Prompt used for LLM-as-a-judge verdict-closure detection.

Prompt for outcome-conditioned rationale elicitation (P12, MIMIC-III, MIMIC-IV, eICU)

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given features of a patient, provide possible clinical rationale for why the patient could {outcome_phrase} during this ICU stay.
Assume the outcome is {outcome_phrase} and list only supporting features.
- Use only the provided features; do not invent data.
- Do not describe normal or healthy findings as abnormal.
- Do not describe abnormal findings as normal or healthy.
- Do not mention any label, class, or numeric outcome.
- Do not discuss the opposite outcome.
- If there is no supporting evidence, leave the rationale blank.

Your answer format must be:
```
## Rationale for {outcome_phrase}
[blank or concise bullet points]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 17: Prompt used to elicit outcome-conditioned rationales from a strong LLM on the in-hospital mortality datasets (P12, MIMIC-III, MIMIC-IV, and eICU). The same prompt structure is used for all four datasets; only the feature-list block is replaced with the dataset-specific list. The placeholder {`outcome_phrase`} is instantiated separately as `in-hospital death` or `survival`, yielding one outcome-specific prompt per outcome per patient.

Prompt for outcome-conditioned rationale elicitation (P19)

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given features of a patient, provide possible clinical rationale for why the patient could experience {assumption}.
Assume the outcome is '{assumption}' and list only supporting features.
- Use only the provided features; do not invent data.
- Do not describe normal or healthy findings as abnormal.
- Do not describe abnormal findings as normal or healthy.
- Do not mention any label, class, or numeric outcome.
- Do not discuss the opposite outcome.
- If there is no supporting evidence, leave the rationale blank.

Your answer format must be:
```
## {header}
[blank or concise bullet points]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 18: Prompt used to elicit outcome-conditioned rationales from a strong LLM on P19 (early sepsis prediction). For each candidate outcome, the placeholders {`assumption`} and {`header`} are paired and substituted as either (`sepsis onset within the next 6 hours`, `Rationale for sepsis`) or (`no sepsis onset within the next 6 hours`, `Rationale for no sepsis`).

Main SFT prompt for TRIAGE (P12, MIMIC-III, MIMIC-IV, eICU) — negative-class-first variant

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given feature of a patient, answer the question below.

Question
Will the patient experience in-hospital death during this ICU stay?

Reasoning by the following process:
1. If the patient indeed survives, which of the patient's given features might be the cause?
2. If the patient indeed experiences in-hospital death, which of the patient's given features might be the cause?
3. Make a final decision: '0' for survival, '1' for in-hospital death.

Your answer format must be as follows:
```
## Rationale for survival
[possible justification if patient survives]

## Rationale for in-hospital death
[possible justification if patient experiences in-hospital death]

## Final Decision
[0 (survival) or 1 (in-hospital death); respond by single number only]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 19: Main SFT prompt used for TRIAGE on the in-hospital mortality datasets (P12, MIMIC-III, MIMIC-IV, and eICU; *negative-class-first* variant). The model is supervised to produce two outcome-conditioned rationale blocks followed by a single `## Final Decision`, with the rationale for the negative class (`survival`, label `0`) preceding the rationale for the positive class (`in-hospital death`, label `1`).

Main SFT prompt for TRIAGE (P12, MIMIC-III, MIMIC-IV, eICU) — positive-class-first variant

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given feature of a patient, answer the question below.

Question
Will the patient experience in-hospital death during this ICU stay?

Reasoning by the following process:
1. If the patient indeed experiences in-hospital death, which of the patient's given features might be the cause?
2. If the patient indeed survives, which of the patient's given features might be the cause?
3. Make a final decision: '0' for survival, '1' for in-hospital death.

Your answer format must be as follows:
```
## Rationale for in-hospital death
[possible justification if patient experiences in-hospital death]

## Rationale for survival
[possible justification if patient survives]

## Final Decision
[0 (survival) or 1 (in-hospital death); respond by single number only]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 20: Main SFT prompt used for TRIAGE on the in-hospital mortality datasets (P12, MIMIC-III, MIMIC-IV, and eICU; *positive-class-first* variant): the rationale for the positive class (`in-hospital death`, label `1`) precedes the rationale for the negative class (`survival`, label `0`). Compared with the negative-class-first variant (Figure 19), only the ordering of the reasoning steps and rationale blocks differs; all other text is identical.

Main SFT prompt for TRIAGE (P19) — negative-class-first variant

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given feature of a patient, answer the question below.

Question
Will the patient experience sepsis onset within the next 6 hours?

Reasoning by the following process:
1. If the patient indeed does not experience sepsis onset within the next 6 hours, which of the patient's given features might be the cause?
2. If the patient indeed experiences sepsis onset within the next 6 hours, which of the patient's given features might be the cause?
3. Make a final decision: '0' for no sepsis onset within the next 6 hours, '1' for sepsis onset within the next 6 hours.

Your answer format must be as follows:
```
## Rationale for no sepsis
[possible justification if patient does not experience sepsis onset within the next 6 hours]

## Rationale for sepsis
[possible justification if patient experiences sepsis onset within the next 6 hours]

## Final Decision
[0 (for no sepsis onset within the next 6 hours) or 1 (for sepsis onset within the next 6 hours); respond by single number only]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 21: Main SFT prompt used for TRIAGE on P19 (*negative-class-first* variant). The model is supervised to produce two outcome-conditioned rationale blocks followed by a single `## Final Decision`, with the rationale for the negative class (`no sepsis`, label `0`) preceding the rationale for the positive class (`sepsis`, label `1`).

Main SFT prompt for TRIAGE (P19) — positive-class-first variant

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given feature of a patient, answer the question below.

Question
Will the patient experience sepsis onset within the next 6 hours?

Reasoning by the following process:
1. If the patient indeed experiences sepsis onset within the next 6 hours, which of the patient's given features might be the cause?
2. If the patient indeed does not experience sepsis onset within the next 6 hours, which of the patient's given features might be the cause?
3. Make a final decision: '0' for no sepsis onset within the next 6 hours, '1' for sepsis onset within the next 6 hours.

Your answer format must be as follows:
```
## Rationale for sepsis
[possible justification if patient experiences sepsis onset within the next 6 hours]

## Rationale for no sepsis
[possible justification if patient does not experience sepsis onset within the next 6 hours]

## Final Decision
[0 (for no sepsis onset within the next 6 hours) or 1 (for sepsis onset within the next 6 hours); respond by single number only]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 22: Main SFT prompt used for TRIAGE on P19 (*positive-class-first* variant): the rationale for the positive class (`sepsis`, label `1`) precedes the rationale for the negative class (`no sepsis`, label `0`). Compared with the negative-class-first variant (Figure 21), only the ordering of the reasoning steps and rationale blocks differs; all other text is identical.

Prompt for one-sided rationale SFT (P12)

I will provide you with medical information from Intensive Care Unit (ICU) visit of a patient, each characterized by number of features.
The list of features are as follows:
{FEATURE_INFO}

Based on the given feature of a patient, answer the question below.

Question
Will the patient experience in-hospital death during this ICU stay?

Reasoning by the following process:
1. Describe the clinical evidence observed in the patient's features.
2. Make a final decision: '0' for survival, '1' for in-hospital death.

Your answer format must be as follows:
```
## Rationale
[clinical evidence observed in the patient's features]

## Final Decision
[0 (survival) or 1 (in-hospital death); respond by single number only]
```

Feature of the patient
{PATIENT_FEATURES}

Figure 23: Prompt template used for the *one-sided rationale SFT* baseline in our reasoning-structure ablation (Table 5), conducted on P12. The supervision retains only the rationale steps that support the ground-truth label from our collected outcome-conditioned traces, so a single `## Rationale` block is followed by the corresponding `## Final Decision`.

Faithfulness: Prompt for locating the input evidence of selected rationale sentences

You are a clinical data annotator for a faithfulness experiment on ICU mortality prediction.

You will be given:
1. Observed patient information from the first 48 hours of ICU stay.
2. Sentences selected from a rationale that a prediction model wrote, based on this patient information, to argue for the outcome "{OUTCOME}".

Your task is to list every piece of data in the patient information that the selected sentences rely on.
The listed data will be erased from the patient information; after the erasure, none of the selected sentences should be derivable from what remains.
Do not judge whether the sentences or the prediction are medically correct.

What to list

For each selected sentence, find every feature it draws on, including features mentioned only as context (e.g., "on mechanical ventilation"), and list the data it relies on:
- Whole-series statements: if the sentence characterizes a feature as a whole (its overall level or range, a trend, a first-to-last change, persistence, or wording such as "throughout", "mostly", "consistently", "repeatedly"), the whole feature is to be erased: use "all".
- Specific statements: if the sentence refers only to particular values, time points, or episodes (e.g., "initial K 6.4", "MAP 56-60 mmHg around 37.5-37.7 h"), give the condition that covers them and any other measurement from which the same statement could still be made: a range of values, a time window, or both. A script erases every pair of the feature that satisfies the condition, so do not copy pairs.
- Invasive and non-invasive blood pressure are different features (SysABP, DiasABP, MAP are invasive; NISysABP, NIDiasABP, NIMAP are non-invasive). If the sentence names one of them, by its feature name or by the word "invasive" or "non-invasive", list only that feature. If the sentence does not say which one it means (e.g., "blood pressure", "diastolic pressure"), apply the rules above to every observed variant. If the feature the sentence names has no measurements for this patient, list the observed variant that the statement evidently describes.
- Clinical concepts and derived quantities: if the sentence names a concept or a computed quantity rather than a feature (e.g., "oliguria", "lactic acidosis", "PaO2/FiO2 ratio"), list the data of the features that directly express or compute it.
- Static descriptors: list age, gender, height, weight, or ICU type (given in the sentence that begins with "A patient is") only if the selected sentence explicitly refers to that descriptor (e.g., "advanced age", "cardiac surgery unit"). Generic phrases such as "critical illness" or "ICU stay" do not refer to the ICU type.
- Cited values may be approximate or rounded; choose bounds that include them (e.g., "in the 50s-60s" is the value range 50 to 69.9). When unsure whether a measurement contributed to a selected sentence, include it.
- Do not list data that is only clinically related to a sentence but not needed to write it.
- If a sentence relies on nothing that is present in the patient information (e.g., it cites unrecorded data or the absence of a measurement), list nothing for it.

Output Format

Return one JSON object and nothing else:
{"deletions": [{"sentence": <sentence id>, "feature": "<feature header, or Age / Gender / Height / Weight / ICUType>", "removed": "all" or {"value": [low, high], "time": [start, end]}}]}

- "removed": use "all" when every pair of the feature is to be erased, and always for a static descriptor. Otherwise give a condition: "value" is a range of measured values and "time" a window in hours since ICU admission. Give either one or both (then both must hold). Bounds are inclusive, null leaves an end open, and a single value or time point is a range with equal bounds, e.g., {"value": [120, null]}, {"time": [0, 6]}, {"value": [6.4, 6.4], "time": [1.5, 1.5]}.
- One object per (sentence, feature) pair. If a sentence needs several separate conditions on one feature (e.g., two episodes), use one object per condition.
- Every condition must select at least one measurement that is present in the patient information.
- Every selected sentence must appear at least once. If nothing is to be erased for a sentence, use {"sentence": <sentence id>, "feature": null, "removed": "none"}.
- Write compact JSON on a single line, without indentation or line breaks.

Input

[Patient Information]
"""{PATIENT_FEATURES}"""

[Selected Sentences]
"""{SELECTED_SENTENCES}"""

Figure 24: Prompt used to locate, in the model input, the evidence behind selected rationale sentences. At deletion ratio $r \in \{0.25, 0.5, 0.75, 1.0\}$, $\lfloor rn \rfloor$ of the n sentences in the rationale block of the predicted outcome are pre-selected at random (seed $42 + i$ for the i-th patient) and passed as {SELECTED_SENTENCES}; {OUTCOME} is survival or in-hospital death.

Faithfulness: Prompt for judging the residual presence of rationale statements

You are an evaluator of clinical rationales written by an ICU mortality prediction model.

You will be given:
1. A numbered list of statements taken from a rationale that the model wrote for the outcome "{OUTCOME}" of one patient.
2. A second rationale for the same outcome and the same patient, which the model wrote later from a modified version of the patient's data.

Your task is to score, for every numbered statement, how completely it has disappeared from the second rationale.
Judge only what is written. Do not judge medical correctness, and do not guess which statements should or should not have changed.

How to compare

For each statement, identify (a) its clinical claim (e.g., "acute kidney injury", "stable blood pressure") and (b) its cited evidence: every feature it names, together with the values, ranges, or trends given for that feature.
Then find the passage of the second rationale that is most similar to the statement and assign exactly one score:

- 5 (Absent): no passage makes this claim or cites this evidence.
- 4 (Revised): a passage makes the same or a closely related claim, but none of the cited evidence is repeated: the features are cited with different numbers, or the claim rests on other features.
- 3 (Finding retained without values): a passage states the same finding for at least one of the cited features (e.g., "creatinine remains normal"), but none of the cited values, ranges, or trend end points is repeated.
- 2 (Partly retained): a passage repeats part of the cited evidence unchanged: at least one feature with the same value, range, or trend, while the rest is missing or different.
- 1 (Retained): a passage makes the same claim and repeats all of the cited evidence: the same features with the same values, ranges, or trends.

Rules

- Abbreviations and synonyms denote the same feature (e.g., "bicarbonate" and "HCO3"). Invasive and non-invasive blood pressure are different features.
- Two numbers are the same if they agree after rounding to the precision used in the statement. A range is the same only if both of its ends agree. A trend is the same if its direction and its end points agree. Words such as "about" or "~" do not matter.
- Wording, order, and interpretation do not matter; only the claim and the cited evidence do.
- If several passages match a statement, use the one that gives the lowest score.
- For scores 1–4, copy the matching passage (one sentence or bullet) verbatim from the second rationale. For score 5, use an empty string.

Output Format

Return one JSON object and nothing else:
{"judgments": [{"statement": <statement id>, "score": 1 or 2 or 3 or 4 or 5, "passage": "<verbatim passage, or empty>"}]}

Every statement must appear exactly once. Write compact JSON on a single line, without indentation or line breaks.

Input

[Statements]
"""{STATEMENTS}"""

[Second Rationale]
"""{SECOND_RATIONALE}"""

Figure 25: Prompt used to score, on a five-point scale (5 = absent, 1 = retained in full), how completely rationale statements disappear after input erasure. {STATEMENTS} lists *all* sentences of the original rationale of the predicted outcome, without marking which ones were targeted by the erasure, and {SECOND_RATIONALE} is the rationale for the same outcome regenerated from the input perturbed at a deletion ratio of 50% (Figure 24); {OUTCOME} is survival or in-hospital death. The judge therefore does not know the label, the predicted probability, the targeted sentences, or the perturbation condition; targeted and untargeted sentences are separated only when the scores are aggregated.

Criteria presented to the clinical experts in the human evaluation

Rate the rationale generated by each method independently along the three criteria below, and record an integer score from 1 to 5 for every criterion. Higher scores indicate higher quality. Ties across methods are permitted: the scores are absolute judgments rather than a ranking. Leave a score blank and annotate the case if it cannot be assessed.

C1. Consistency (1–5 per method)
Definition. Consistency measures how well the generated rationale aligns with the model's prediction. It evaluates whether the reasoning contains contradictions.
- 5 (Excellent): The rationale is entirely consistent with the prediction. No contradictions are present.
- 4 (Good): Minor inconsistencies are present but do not significantly detract from the overall coherence.
- 3 (Fair): Some inconsistencies exist that partially undermine the rationale's reliability.
- 2 (Poor): Significant inconsistencies exist, making the rationale difficult to trust.
- 1 (Very Poor): The rationale is fundamentally contradictory to the prediction.

C2. Correctness (1–5 per method)
Definition. Correctness assesses the medical validity of the knowledge and reasoning presented in the rationale.
- 5 (Excellent): The rationale is entirely medically accurate and reflects evidence-based clinical knowledge.
- 4 (Good): Minor inaccuracies are present but do not affect the overall clinical validity.
- 3 (Fair): The rationale contains some medically incorrect statements that could mislead clinicians.
- 2 (Poor): Multiple inaccuracies significantly reduce the credibility of the rationale.
- 1 (Very Poor): The rationale is largely or entirely medically incorrect, making it unusable.

C3. Helpfulness (1–5 per method)
Definition. Helpfulness measures the extent to which the rationale aids the prediction toward the correct diagnosis.
- 5 (Excellent): The rationale strongly supports the correct prediction and adds valuable clinical insights.
- 4 (Good): The rationale is helpful overall but lacks some critical insights.
- 3 (Fair): The rationale provides some useful guidance but is incomplete or not compelling.
- 2 (Poor): The rationale adds minimal value to the prediction and lacks actionable insights.
- 1 (Very Poor): The rationale is unhelpful and does not contribute meaningfully to the diagnosis.

Optional patient input
The observed patient record may be consulted when needed. Time is reported as hours since ICU admission, and measurements are listed as (Time, Value) in chronological order. If the record is consulted for a case, the same input is applied while rating all four methods. Ground-truth outcomes are not provided to the raters.

Figure 26: Definitions and rating scales presented to the clinical experts in the human evaluation (Appendix F.1), adapted from Jiang et al. (2025a). Each rationale is scored independently on consistency, correctness, and helpfulness using a five-point integer scale, and the raters are not given the ground-truth outcome.

Prompt for the interpretation of the IG result of STraTS.

You are an interpreter for explanations of a trained ICU mortality prediction model.
You will be given a textualized summary of the model's prediction evidence. This summary was generated from important patient-specific observations selected from the model. Your task is to interpret what the provided evidence means about the model's prediction.

Input

List of features
{FEATURE_INFO}
Evidence Summary
{XAI_RESULT_TEXTUALIZE}

Your Task

Explain what the provided evidence summary means about the model's prediction.
Interpretation means:
- identify what kinds of observations the model relied on,
- distinguish evidence for the predicted class from evidence for the opposite class,
- group related observations when appropriate,
- explain the model's prediction in plain language.
Interpretation does NOT mean:
- making a new clinical judgment,
- deciding whether the model is medically correct,
- inferring missing diagnoses, treatments, or outcomes,
- explaining the XAI method,
- discussing attribution scores or importance scores.

Rules

- Use only the information in the textualized evidence summary.
- Do not introduce new clinical interpretations that are not directly supported by the summary.
- Do not overstate the evidence.
- Do not evaluate the adequacy of the evidence.
- Do not add unnecessary concluding statements.
- Keep the interpretation concise.

Output Format

Only write the interpretation of the evidence predicted by the model. Write a concise interpretation ($\leq$ 10 sentences for each section) along with specific figures in the following format:
Rationale for survival
- Interpretation of evidences
Rationale for in-hospital death
- Interpretation of evidences

Figure 27: Prompt for the interpretation of the IG result of STraTS.

Evaluation: Prompt for IDEA assessment tool (1/4)

You are an expert in clinical reasoning assessment and ICU prognostic reasoning. Your task is to evaluate one clinical reasoning trace and score how medically reasonable it is for predicting in-hospital mortality during the ICU stay.

Input Data

[Clinical Question]
"""Will the patient experience in-hospital death during this ICU stay?"""

[Patient Features]
"""[[PATIENT_FEATURES]]"""

[Reasoning]
"""[[MODEL_REASONING]]"""

Task Instructions

Evaluate whether the reasoning trace is clinically reasonable, faithful to the provided patient features, temporally aware, balanced, and appropriately calibrated for ICU in-hospital mortality prediction.
Your job is to judge the quality of the medical reasoning based only on the provided patient features.
Your job is NOT to determine the true outcome.
Your job is NOT to infer or reward matching any hidden label.
Your job is NOT to reward longer, more fluent, more confident, or more detailed writing unless the added detail is accurate, patient-specific, clinically relevant, and proportional to the evidence.
Score the reasoning using the adapted Revised-IDEA rubric below.

Mandatory Factuality and Data-Use Rules

Use only the provided patient features.
You may use general medical knowledge to interpret the provided features, but you must not add clinical facts that are not present.
Do NOT assume unstated diagnoses, medications, procedures, organ support, code status, goals of care, complications, outcomes, or information after the prediction window.
Do NOT treat missing data as normal or abnormal without justification.
Do NOT infer the true label.
Do NOT reward the reasoning for matching any hidden label.
Do NOT reward a model simply because it gives a more extreme mortality probability.
Do NOT reward generic medical knowledge unless it is accurately connected to this patient.
Penalize reasoning that:
- Hallucinates unsupported major diagnoses, organ failures, treatments, procedures, or events.
- Contradicts the provided patient data.
- Misreads or reverses the temporal trajectory.
- Treats time-series measurements as static when the trend is prognostically important.
- Overemphasizes minor or isolated abnormalities while ignoring the main prognostic problem.
- Gives very high or very low mortality risk without evidence proportional to that confidence.
- Cherry-picks only risk-increasing or only risk-decreasing evidence when the data are mixed.

Adapted Revised-IDEA Rubric for ICU Mortality Reasoning

Score the reasoning out of 10 points using four domains:

I. Interpretive Prognostic Summary: 0-4 points
D. Differential Prognostic Assessment: 0-2 points
E. Explanation of Lead Prognostic Judgment: 0-2 points
A. Alternative Prognostic Explanation / Counterevidence: 0-2 points

The total score is I + D + E + A, from 0 to 10.
The scoring should emphasize clinical reasoning quality, factuality, temporal awareness, and calibration. Do not mechanically reward length or the number of facts listed.

Figure 28: Prompt used for the quantitative evaluation of generated reasoning traces (1/4). We adopt the IDEA assessment tool as the evaluation rubric.

Evaluation: Prompt for IDEA assessment tool (2/4)

I — Interpretive Prognostic Summary, 0-4 points

This assesses whether the reasoning gives an accurate, patient-specific, temporally aware summary of the patient's main prognostic problem.
Prefer reasoning that accurately identifies, when supported by the patient features:

1. Key baseline or contextual mortality-risk factors:
- Examples: age, major comorbidities, chronic organ disease, frailty-relevant context, admission context, or other baseline risks.
2. Main acute ICU problem or organ dysfunction:
- Examples: respiratory failure, shock or hemodynamic instability, renal dysfunction, metabolic acidosis, neurologic depression, hepatic or coagulation dysfunction, infection signal, multi-organ dysfunction, or other acute problems explicitly supported by the data.
3. Illness time course or trajectory:
- Correctly distinguishes worsening, improving, fluctuating, or stable course.
- Gives special attention to persistent instability or late deterioration near the end of the observation window.
4. Clinically meaningful prognostic abstractions:
- Examples: persistent hypotension, escalating oxygen needs, worsening renal function, severe acidosis, multi-organ dysfunction, improving hemodynamics, stable oxygenation, transient isolated abnormality.
- These abstractions must be grounded in the provided data.

Scoring anchors:
- 4: Accurate, patient-specific, and temporally aware summary that captures the main acute problem, key relevant baseline/contextual risks if provided, trajectory, and meaningful prognostic abstractions.
- 3: Mostly accurate summary with the main prognostic problem identified, but with a minor omission, limited temporal interpretation, or limited abstraction.
- 2: Partially accurate summary; mentions some relevant patient-specific factors but misses an important main problem, trajectory, or severity distinction, or includes minor unsupported statements.
- 1: Minimal or generic summary; only one relevant patient-specific element is accurately identified, or the reasoning mostly lists data without integrating the main prognostic problem.
- 0: No meaningful prognostic summary, or the summary is mostly contradicted by the patient features or dominated by hallucinated information.
Do not penalize the reasoning for not mentioning information that is not provided.
Do penalize it for inventing missing baseline risks, diagnoses, treatments, or organ failures.

D — Differential Prognostic Assessment, 0-2 points

This assesses whether the reasoning considers competing prognostic interpretations and explicitly prioritizes the most plausible mortality-risk interpretation.
For this ICU mortality task, "differential" means competing mortality-risk hypotheses, not a diagnostic differential diagnosis.

Relevant competing interpretations include:
- High vs moderate vs low mortality risk.
- Mortality risk driven by multi-organ dysfunction vs lower risk because abnormalities are mild or transient.
- High-risk deteriorating trajectory vs stabilization or survival trajectory.
- Risk-increasing signals vs risk-decreasing signals.
- Uncertainty due to mixed or incomplete evidence.

Scoring anchors:
- 2: Clearly states a lead mortality-risk interpretation and explicitly discusses at least one plausible alternative or countervailing interpretation, then prioritizes the lead interpretation using patient-specific data.
- 1: Gives a lead mortality-risk interpretation but alternatives are implicit, generic, weakly discussed, or not clearly prioritized; or lists competing ideas without clearly weighing them.
- 0: No meaningful prognostic assessment; only a bare conclusion, generic statement, or diagnostic laundry list that does not support mortality-risk reasoning.
If the evidence is strongly one-directional, the reasoning can still receive 2 points if it explicitly explains why alternative interpretations are less supported and acknowledges remaining uncertainty appropriately.

Figure 29: Prompt used for the quantitative evaluation of generated reasoning traces (2/4). We adopt the IDEA assessment tool as the evaluation rubric.

Evaluation: Prompt for IDEA assessment tool (3/4)

E — Explanation of Lead Prognostic Judgment, 0-2 points

This assesses whether the reasoning explains why its lead mortality-risk judgment follows from the objective patient data.

Prefer reasoning that:
- Clearly links objective patient-specific evidence to the lead mortality-risk judgment.
- Uses clinically relevant data such as vital signs, laboratory values, organ-support indicators, severity scores, comorbidities, or temporal trends when provided.
- Interprets the severity and persistence of abnormalities appropriately.
- Gives a final risk interpretation that is proportional to the evidence.

Calibration requirements:
- Very high mortality risk requires strong evidence, such as severe or persistent instability, major organ dysfunction, multiple converging high-risk features, or marked late deterioration.
- Very low mortality risk should not ignore major organ dysfunction, severe derangements, or sustained deterioration.
- Mild, isolated, or transient abnormalities should not be overstated as extreme risk.
- If evidence is mixed, the reasoning should express appropriate uncertainty.

Scoring anchors:
- 2: Uses at least two objective patient-specific data points or temporal trends to support the lead judgment, interprets them clinically and accurately, and gives a calibrated final mortality-risk interpretation.
- 1: Provides some patient-specific support, but the explanation is incomplete, weakly linked, based on only one main data point, partially miscalibrated, or contains minor unsupported claims.
- 0: No explanation of the lead mortality-risk judgment, or the explanation is unsupported, contradicted by the data, dominated by hallucinated facts, or purely generic.

Do not give full credit for simply naming abnormal values without explaining their prognostic relevance.

A — Alternative Prognostic Explanation / Counterevidence, 0-2 points

This assesses whether the reasoning identifies and explains patient-specific evidence that could support a different, lower-risk, higher-risk, or more uncertain interpretation.
For this ICU mortality task, "alternative diagnosis explained" is adapted to mean alternative prognostic explanation or counterevidence.

Relevant counterevidence may include:
- Reassuring or risk-decreasing factors.
- Stable or improving organ function.
- Transient rather than sustained abnormalities.
- Lack of support for severe conditions that the reasoning might otherwise imply.
- Risk-increasing evidence that contradicts an overly reassuring conclusion.
- Mixed evidence requiring uncertainty.
- Absence of provided evidence for major interventions or complications, without treating missing data as definitely normal.

Scoring anchors:
- 2: Identifies and explains at least two patient-specific countervailing data points, alternative prognostic explanations, or sources of uncertainty, and weighs them fairly against the lead judgment.
- 1: Mentions one patient-specific countervailing data point or gives a generic/limited discussion of uncertainty or alternative risk interpretation.
- 0: No meaningful counterevidence or alternative prognostic explanation; cherry-picks only evidence supporting the conclusion; or treats missing information as definitely normal or definitely abnormal.

Counterevidence can point in either direction:
- If the lead judgment is high mortality risk, counterevidence may include stabilization, improvement, preserved organ function, or transient abnormalities.
- If the lead judgment is low mortality risk, counterevidence may include persistent instability, organ dysfunction, severe derangements, or late deterioration.

Figure 30: Prompt used for the quantitative evaluation of generated reasoning traces (3/4). We adopt the IDEA assessment tool as the evaluation rubric.

Evaluation: Prompt for IDEA assessment tool (4/4)

Overall Score Calibration

Use the following general score interpretation:

- 9-10: Excellent reasoning. Accurate, patient-specific, temporally aware, balanced, well-explained, and well-calibrated. No important hallucinations or major omissions.
- 7-8: Good reasoning. Mostly accurate and clinically grounded, with minor omissions, limited counterevidence, or mild calibration issues.
- 5-6: Mixed reasoning. Some accurate patient-specific evidence, but important omissions, weak temporal reasoning, one-sided assessment, or moderate calibration problems.
- 3-4: Weak reasoning. Generic, poorly grounded, several unsupported claims, major omissions, or significant misreading of the clinical picture.
- 0-2: Poor reasoning. Little to no patient-specific reasoning, severe hallucinations, major contradictions, hidden-label reasoning, or conclusion unsupported by the provided data.

Internal Deliberation

Before producing the final judgment, internally check:
1. What objective patient-specific evidence the reasoning used correctly.
2. Any unsupported claims, hallucinations, contradictions, or missing-data assumptions.
3. Whether the reasoning correctly handled the temporal trajectory.
4. Whether the reasoning considered counterevidence or uncertainty.
Do not output a long chain-of-thought. Output only the JSON object below.

Output Format

Return valid JSON only. Do not include markdown, comments, or extra text.
The score and all subscores must be integers. The final score must equal the sum of the four subscores after applying any cap or penalty.
The confidence field reflects your confidence in the reasoning-quality score, not confidence in the patient's mortality outcome.
The confidence field must be exactly one of: "High", "Medium", or "Low".
Use empty arrays if there are no notable items.

Return this JSON object:

```
{
  "score": 0 - 10,
  "confidence": "High" — "Medium" — "Low",
  "subscores": {
     "I_interpretive_prognostic_summary": 0,
     "D_differential_prognostic_assessment": 0,
     "E_explanation_of_lead_prognostic_judgment": 0,
     "A_alternative_or_counterevidence_explained": 0
  },
  "rationale": "Briefly explain the score. Mention the most important patient-specific evidence, temporal reasoning, calibration, counterevidence, and any major errors or cap applied.",
  "key_strengths": [],
  "key_errors": []
}
```

Figure 31: Prompt used for the quantitative evaluation of generated reasoning traces (4/4). We adopt the IDEA assessment tool as the evaluation rubric.

Prompt used to detect severe hallucination in reasoning traces

You are a clinical evidence-grounding judge for ICU mortality/survival rationales.
You will be given:
1. Observed patient information from the first 48 hours of ICU stay.
2. A rationale explaining why the patient is predicted to die or survive.
Your task is to judge whether the rationale is grounded in the observed patient information, or whether it contains serious hallucination: concrete clinical facts that are completely absent from the patient data.
Return `True` if the rationale's substantive clinical evidence is present in, paraphrased from, summarized from, or reasonably inferred from the patient information.
Return `False` if the rationale uses at least one clinically meaningful fabricated fact as evidence, such as a diagnosis, treatment, event, demographic detail, lab/vital value, trend, medication, procedure, or condition that is not present in the patient information or is clearly contradicted by it.
Be lenient. The goal is to detect serious hallucination, not minor imprecision.
Do not mark `False` for small wording issues, approximate values, rounding, broad ranges, mild exaggeration, or imperfect temporal alignment.
If a value or feature appears anywhere in the patient data and the rationale refers to it in a broadly consistent way, consider it grounded.
For example, if the patient data contains PaO2 values of 98, 100, and 125 mmHg, then a rationale saying "PaO2 around 98–125 mmHg indicating adequate oxygenation" should be considered grounded, even if the exact timing is not perfectly aligned with another feature such as mechanical ventilation.
Do not require every cited feature to co-occur at the exact same timestamp unless the rationale makes a very specific timing claim that is central and clearly contradicted.
Accept clinical interpretations such as "stable," "adequate oxygenation," "renal dysfunction," "shock," or "high mortality risk" if they are reasonably supported by the observed measurements.
Do not evaluate whether the mortality/survival prediction is medically correct; only judge whether the rationale is based on information actually observed in the input.
Mark `False` only when there is a clear, material unsupported claim, for example:
- The rationale cites lactate, creatinine, vasopressors, mechanical ventilation, sepsis, pneumonia, cancer, renal failure, surgery, or another condition/treatment/event that is not present anywhere in the patient data.
- The rationale gives a concrete numeric value or range that is grossly inconsistent with the observed values.
- The rationale describes a trend that is clearly opposite to the observed trend.
- The rationale relies on medical history or outcome information not included in the patient data.
If the unsupported statement is only a minor non-evidentiary phrase and the main clinical evidence is grounded, return `True`.
When uncertain whether something is a minor overstatement or a serious hallucination, prefer `True`.

Input:

```
[Patient Information]
"""{patient_information}"""
[Rationale]
"""{rationale}"""
```

Output only a Python-style dict with exactly these two keys:

```
{
 "reason": "<brief explanation in 10 sentences or fewer>",
 "result": True or False
}
```

Figure 32: Prompt used to detect severe hallucination in reasoning traces. The judge model decides whether the rationale's substantive clinical evidence is grounded in the input. Only clearly fabricated facts such as unobserved diagnoses, treatments, lab values, or trends contradicting the data are marked as hallucinations.