EARLY WARNING FOR FIRST FAILURE IN MULTI-ROUND LLM GENERATION VIA SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly operate through sequential generation, including multi-turn interactions and multi-step reasoning. Longer reasoning chains can improve performance but also create more opportunities for errors and hallucinations. Consistency-based uncertainty estimates can help assess reliability, yet an early error may cause subsequent outputs to remain internally consistent with a corrupted trajectory, yielding deceptively low uncertainty. This motivates anticipating the *first failure* along a generation trajectory. To tackle this problem, we formulate first-failure monitoring as a discrete-time survival problem, scoring each step before its output is released. Beyond the process formulation, we further propose propagated historical instability to capture pre-failure dynamics, and apply a conformal calibration layer that bounds, with finite-sample validity, the probability that a first failure occurs without a warning. We further show that access to propagated historical instability cannot worsen the oracle warning-burden frontier by score-class nesting, and derive conditions under which this additional history yields a strict burden reduction. Empirically, we evaluate the framework across four HalluHard domains with two open LLMs and on PRM800K reasoning chains, comparing against raw uncertainty scores, alternative feature sets, and two survival models. Within our hazard model, adding propagated historical instability improves both AUROC and AUPR in all 12 settings, and the calibrated propagation-aware rule achieves the lowest warning burden among admissible constructions across the evaluated operating points.

1 INTRODUCTION

Modern LLM systems increasingly operate through multi-step generation. A user task may unfold over many conversational turns, while a single answer may itself emerge through a long sequence of reasoning steps. In chain-of-thought reasoning, later steps similarly depend on the preceding reasoning trajectory. Yet, generation quality can degrade as multi-turn interactions become longer (Laban et al., 2026). Recent studies suggest that this degradation may arise in part because subsequent generations condition on the accumulated interaction history, allowing early errors to propagate through later steps and become difficult to recover from (Laban et al., 2026; Chenkang et al., 2026). Therefore, anticipating when the first failure occurs is particularly important in such sequential settings, and detecting failures after they occur is not enough. This motivates our central research question:

> *Research Question: Given a sequential generation process that has remained correct so far, how likely is the next step to be the first failure?*

To answer this question, a natural approach is to monitor the evolving interaction using existing uncertainty-quantification methods (Kuhn et al., 2023; Zhang et al., 2026b) or confidence-estimation methods (Zhang et al., 2026a). However, most such methods attach uncertainty to the generation at the current step, with the primary goal of assessing the reliability of that response. When applied to a long multi-turn interaction, a reliability score for each generated response may not be sufficient. Once an earlier failure occurs, the interaction enters an erroneous trajectory, and later generations may remain internally consistent with the corrupted context. They may therefore appear reliable even when they are wrong, making post-failure uncertainty unsuitable for first-failure monitoring.

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.

For first-failure monitoring, we therefore consider only interactions that have remained failure-free up to the current step, which we refer to as the still-at-risk process. Moreover, by definition, the score for assessing the risk that the current step will be the first failure must be available before the monitored output is released. This makes early warning more demanding than post-hoc evaluation, where the generated response is already available for inspection. Although these uncertainty quantities may provide useful signals, they are not themselves probabilities of first failure. For example, an uncertain response does not necessarily mark the first failure. Conversely, a response with low uncertainty may still be the first failure. Hence, we first summarize the information available before the monitored output is released. Then, we learn how these summaries relate to first failure among states that remain failure-free.

To align the statistical target with the monitoring task, we formulate first-failure monitoring in sequential LLM generation as a survival problem, which naturally captures key features of our setting. For example, an interaction or reasoning trajectory that ends without an observed failure can be treated as right-censored, while the discrete-time hazard represents the probability that the current step is the first failure, conditional on no earlier failure having occurred. Although the discrete-time hazard can be estimated using binary regression, the survival formulation determines the appropriate risk set and censoring structure used for estimation. We then estimate this first-failure risk from pre-response uncertainty summaries and other covariates using a discrete-time hazard model.

However, building this hazard also raises a difficulty in identifying informative pre-response signals. A natural starting point is current-step semantic uncertainty, which measures disagreement in meaning across alternative generations from the same history. Yet, this signal describes only the current state. In particular, ignoring information from earlier steps may discard useful signals about first-failure risk. Hence, a natural question is how to incorporate historical information into the risk estimate. The challenge is that, under the still-at-risk process, this historical information cannot come from observed past errors, since every preceding step is failure-free by construction. To address this challenge, we model pre-failure instability as the historical signal. For example, two conversations with similar current uncertainty carry different first-failure risks if one has exhibited persistent instability over earlier steps while the other has remained stable. This motivates augmenting current semantic uncertainty with a historical instability feature that summarizes pre-failure information over recent steps.

Finally, a hazard score by itself does not determine when a warning should be issued. We therefore calibrate the warning threshold using conformal risk control (Angelopoulos et al., 2024), which yields a distribution-free bound on the marginal probability that a first failure goes unwarned. The construction applies to both multi-turn conversations and multi-step reasoning, with the sequential unit being a conversation turn or a reasoning step, respectively.

Contributions. In this paper, we make four main contributions. **First**, we formulate prospective first-failure monitoring in sequential LLM generation as a discrete-time survival problem, explicitly defining the at-risk process, censoring structure, and pre-response first-failure hazard for both multi-turn dialogue and chain-of-thought reasoning. **Second**, we introduce a propagation-aware hazard score that combines current semantic uncertainty with historical pre-failure instability. For the resulting warning rule, we provide a finite-sample statistical guarantee that controls the probability that a first failure occurs without a warning. **Third**, we theoretically analyze warning efficiency, measured by warning rate and unnecessary-warning rate, under a constraint on missed first failures. We show that incorporating historical information cannot worsen the optimal warning burden, and derive conditions under which it leads to a strict reduction in warning burden for the fitted rule. **Fourth**, we empirically demonstrate improved predictive performance and reduced warning burden across HalluHard and

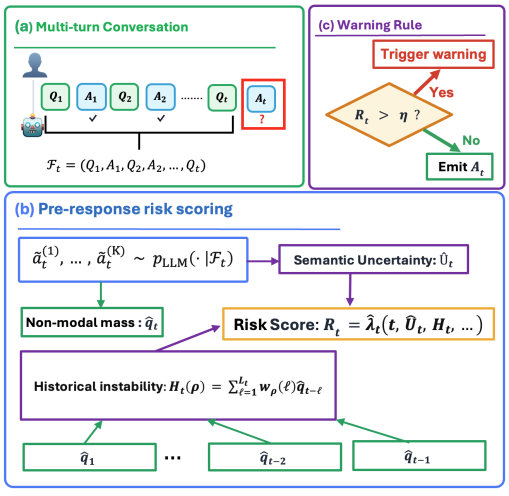

Figure 1: Overview of the proposed early-warning framework and methodology.

PRM800K. Adding the historical feature to our discrete-time hazard model improves both AUROC and AUPR in all 12 evaluated settings, while the calibrated full rule attains the lowest warning burden among admissible constructions across the evaluated operating points.

2 Related Works

Our work connects three strands of literature: sequential reliability and verification in LLMs, uncertainty and confidence estimation for language generation, and survival modeling.

Sequential reliability and verification in LLMs. Recent work has increasingly studied reliability beyond isolated single-turn generation. At the dialogue level, multi-turn evaluations show that LLM performance can deteriorate over extended interactions and that early mistakes may persist through later turns (Laban et al., 2026). Beyond this broad degradation, recent work has identified more specific and measurable failure patterns, such as position-weighted inconsistency across follow-up interactions, which places greater emphasis on earlier-turn behavior (Li et al., 2025), and the loss of claim-level factual support as a dialogue proceeds (Lewis et al., 2026). At the reasoning level, process-supervision methods evaluate intermediate steps after they have been generated and provide step-level feedback on reasoning quality (Lightman et al., 2024), while recent verification benchmarks study the localization of the first erroneous step within a generated reasoning trajectory (Pandit et al., 2026). Although these works establish the importance of monitoring sequential generation, they do not provide a prospective signal for first-failure risk among trajectories that have remained failure-free so far.

Uncertainty and confidence estimation in LLMs. Beyond characterizing and modeling these sequential failure phenomena, a complementary line of work estimates uncertainty in LLM generations as a signal of response reliability. Black-box approaches have used disagreement across sampled generations to detect factual unreliability (Manakul et al., 2023; Xiong et al., 2024), while semantic uncertainty methods account for the fact that lexically different generations may express the same meaning by aggregating samples into semantic equivalence classes (Kuhn et al., 2023; Farquhar et al., 2024). These methods primarily quantify uncertainty at a single state, not how it evolves along a trajectory. Recent studies have therefore extended uncertainty-estimation methods to multi-turn interactions, where confidence is conditioned on the accumulated conversation history and evaluated through turn-wise calibration and confidence dynamics (Zhang et al., 2026b;a). These developments show that history can be important for uncertainty estimation. However, response-level uncertainty is informative but not sufficient to determine when the process should be interrupted.

Survival modeling for sequential LLM. Survival modeling provides a natural framework for the timing of a first event under censoring, with discrete-time hazard models allowing the event risk to evolve with sequentially observed covariates (Singer & Willett, 1993; Cox, 1972). Recently, it has also been used to characterize LLM behavior in sequential settings. Li et al. (2026) formulate the first inconsistent response in multi-turn adversarial interactions as a time-to-event outcome and evaluate several survival models using semantic-drift features. Topol (2026) model the time to a successful jailbreak across repeated trials of the same adversarial prompt. *Our setting differs in both the event of interest and the monitoring objective.* We target the first factual or reasoning failure and require the risk score to be computable before the monitored output is released. We further incorporate propagated pre-failure instability into the risk score and calibrate the resulting warning rule to provide finite-sample control of the probability that a first failure goes unwarned. The same formulation covers both multi-turn dialogue and multi-step reasoning, including settings in which the monitoring horizon may end before the reasoning trajectory itself.

3 Methodology

Problem Formulation. In this study, we consider two sequential LLM generation settings: multi-turn conversations and chain-of-thought reasoning. We first present the problem formulation for the multi-turn conversation setting. Let conversation i be represented by $\mathcal{T}_i = \{(x_{i,1}, a_{i,1}), \ldots, (x_{i,T_i}, a_{i,T_i})\}$, where $x_{i,t}$ is the user input at turn t, $a_{i,t}$ is the LLM response, and T_i is the observed conversation length. For prediction at turn t, define the pre-response history $\mathcal{F}_{i,t} = (x_{i,1}, a_{i,1}, \ldots, x_{i,t-1}, a_{i,t-1}, x_{i,t})$. This is the information available before the model

releases the response $a_{i,t}$. Then, we define the factual failure indicator

$$E_{i,t} = \mathbf{1}\{a_{i,t} \text{ contains at least one factual error or factually misleading claim}\}.$$

The *first factual failure time* is $\tau_i = \inf\{t : E_{i,t} = 1\}$ with the convention $\inf \varnothing = \infty$. Since a failure may not occur within the observed horizon T_i, we define the censored failure-time pair $Y_i = \min\{\tau_i, T_i\}$ and $\Delta_i = \mathbf{1}\{\tau_i \le T_i\}$. $\Delta_i = 1$ indicates that the first factual failure is observed within the conversation, whereas $\Delta_i = 0$ indicates right-censoring at the observed conversation length T_i. We note that all risk sets and calibration losses in later sections are defined with respect to the observed conversational process. The at-risk set is $\mathcal{R} = \{(i,t) : 1 \le t \le Y_i\}$. Because $t \le Y_i = \min\{\tau_i, T_i\}$ implies $t \le \tau_i$, membership $(i,t) \in \mathcal{R}$ means that conversation i has not factually failed before turn t. For each at-risk pair $(i,t) \in \mathcal{R}$, define the event label $Z_{i,t} = \mathbf{1}\{Y_i = t, \Delta_i = 1\}$, which equals $E_{i,t}$. The main uncertainty target is the discrete-time first-failure hazard

$$\lambda_{i,t} = \mathbb{P}(\tau_i = t \mid \tau_i \ge t, T_i \ge t, \mathcal{F}_{i,t}) = \mathbb{P}(E_{i,t} = 1 \mid E_{i,1} = \cdots = E_{i,t-1} = 0, T_i \ge t, \mathcal{F}_{i,t}). \quad (1)$$

This quantity measures the probability of a factual failure at turn t, conditional on the conversation having reached turn t without a prior factual failure. This is the central UQ target of the proposed framework. The analogous formulation for multi-step reasoning is deferred to Appendix A.1.

Implication for propagation feature modeling. In our formulation, once the first failure occurs, all subsequent post-failure states are removed from the risk set. Consequently, for any $(i,t) \in \mathcal{R}$, $E_{i,s} = 0$, for $s < t$. Thus, within the at-risk process, "past failures" cannot be directly propagated as observed covariates. The appropriate object to propagate is instead a pre-failure latent instability measure, estimated from the conversation state before failure occurs. This observation motivates the semantic instability scores and the propagation kernel introduced in the following section.

Propagation of historical instability. For a fixed pre-response history $\mathcal{F}_{i,t}$, an LLM can be queried repeatedly before the final response is released, providing multiple diagnostic responses from the same pre-response state. We exploit this generative structure to estimate semantic uncertainty (Kuhn et al., 2023). Specifically, for each at-risk state $(i,t) \in \mathcal{R}$, we draw K diagnostic responses and group them into semantic equivalence classes with empirical proportions $\widehat{p}_{i,t,c}$. We use the Miller–Madow-corrected semantic entropy $\widehat{U}_{i,t}$ as the current-turn uncertainty feature and define the non-modal semantic mass $\widehat{q}_{i,t} = 1 - \max_c \widehat{p}_{i,t,c}$, which measures the mass outside the modal semantic class. Full details of the diagnostic sampling, semantic grouping, and entropy construction are provided in Appendix A.2. Within the at-risk set, since no observed factual failure has occurred before turn t, past factual failures cannot be propagated as observed covariates. Thus, we model the propagation of latent semantic instability. To motivate a parsimonious history kernel, consider a working state-space model in which latent instability follows approximately AR(1), $\xi_{i,t} = \phi\xi_{i,t-1} + \epsilon_{i,t}$ where $0 < \phi < 1$ and $\epsilon_{i,t}$ is a mean-zero innovation, and is observed through a noisy proxy, $\widehat{q}_{i,t} = \xi_{i,t} + \nu_{i,t}$, where $\nu_{i,t}$ is measurement noise. Because the latent state $\xi_{i,t}$ is not directly observed, prediction from the observed proxy history generally depends on multiple past values of $\widehat{q}_{i,t}$. Under a standard scalar linear state-space approximation, the corresponding steady-state linear filter assigns geometrically decaying influence to past proxy observations, with an effective decay rate determined jointly by the latent persistence and the process and measurement noise levels. Motivated by this filtering structure, we summarize historical instability using a finite-memory normalized geometric average. For a fixed memory length L, define $L_t = \min\{L, t-1\}$. For $t = 1$, we set $H_{i,1}(\rho) = 0$. For $t \ge 2$, define

$$H_{i,t}(\rho) = \sum_{\ell=1}^{L_t} w_\rho(\ell)\widehat{q}_{i,t-\ell}, \quad (2)$$

where $w_\rho(\ell) \ge 0$, $\sum_{\ell=1}^{L_t} w_\rho(\ell) = 1$, and ρ is an effective historical decay parameter. Thus $H_{i,t}(\rho) \in [0,1]$. In this work, we use the normalized geometric kernel, that is $w_\rho(\ell) = \frac{(1-\rho)\rho^{\ell-1}}{1-\rho^{L_t}}$, where $0 < \rho < 1$. Smaller values of ρ place more weight on recent turns, whereas as $\rho \uparrow 1$, the weights approach the uniform kernel, $w_\rho(\ell) \to 1/L_t$.

Propagation-Aware Hazard Scoring. Given the current semantic dispersion score $\widehat{U}_{i,t}$ and the historical instability score $H_{i,t}(\widehat{\rho})$, we model the discrete-time first-failure hazard as

$$\operatorname{logit}\{\lambda_\theta(i,t)\} = \beta_0 + f(t;\gamma) + \beta_1\widehat{U}_{i,t} + \beta_2 H_{i,t}(\widehat{\rho}) + \beta_3^\top B_{i,t}, \quad (3)$$

where $f(t;\gamma) = \sum_{k=1}^{d_f} \gamma_k \psi_k(t)$. Here $\{\psi_k\}_{k=1}^{d_f}$ is a fixed basis over the turn index, such as low-order polynomials or spline basis functions. To ensure identifiability of the intercept β_0, the constant function is excluded from the basis. The parameter β_1 captures current semantic dispersion, whereas β_2 captures propagated historical instability. The covariate vector $B_{i,t}$ contains additional pre-response features, such as context length. Equivalently,

$$\lambda_\theta(i,t) = \sigma\left\{\beta_0 + f(t;\gamma) + \beta_1 \widehat{U}_{i,t} + \beta_2 H_{i,t}(\widehat{\rho}) + \beta_3^\top B_{i,t}\right\}, \qquad \sigma(u) = \frac{1}{1+\exp(-u)}.$$

The kernel parameter $\widehat{\rho}$ is selected on the training split by grid search over a finite grid Φ fixed in advance. For each candidate $\rho \in \Phi$, we construct $H_{i,t}(\rho)$, fit the hazard model, and retain the value maximizing the likelihood on an internal validation subset of the training data. The resulting score-construction procedure is fixed before conformal calibration. Hence, the hazard parameters $\theta = (\beta_0, \gamma, \beta_1, \beta_2, \beta_3)$ are estimated on the training at-risk set by maximizing the log-likelihood

$$\ell(\theta) = \sum_{(i,t)\in\mathcal{R}_{\text{train}}} \left[Z_{i,t} \log \lambda_\theta(i,t) + (1 - Z_{i,t}) \log\{1 - \lambda_\theta(i,t)\}\right].$$

In practice, we use penalized likelihood, $\widehat{\theta} = \arg\min_\theta \{-\ell(\theta) + \Omega(\theta)\}$, where $\Omega(\theta)$ is a regularization term for finite-sample stability. The final pre-response risk score is $R_{i,t} = \widehat{\lambda}_{i,t} = \lambda_{\widehat{\theta}}(i,t)$.

Calibrated Warning Thresholds. Given the estimated hazard score, we trigger a warning whenever $R_{i,t} > \eta$. Because $R_{i,t}$ is computed before releasing $a_{i,t}$, a warning issued at the failure turn $t = \tau_i$ is considered successful. For a fixed threshold η, define the conversation-level missed-warning loss, $L_i(\eta) = \Delta_i \mathbf{1}\{\max_{1\le t\le Y_i} R_{i,t} \le \eta\}$. Thus, $L_i(\eta) = 1$ when an observed first failure occurs without a warning at or before that failure. Since $L_i(\eta)$ is nondecreasing in η, we calibrate the largest threshold satisfying a finite-sample corrected empirical-risk constraint. Let $\widehat{R}_{\text{cal}}(\eta) = \frac{1}{n_{\text{cal}}} \sum_{i\in\mathcal{I}_{\text{cal}}} L_i(\eta)$, where the score-construction pipeline is fitted using only the training set and then held fixed during calibration. Over a finite threshold grid Γ, we choose

$$\widehat{\eta} = \max\left\{\eta \in \Gamma : \frac{n_{\text{cal}}}{n_{\text{cal}}+1} \widehat{R}_{\text{cal}}(\eta) + \frac{1}{n_{\text{cal}}+1} \le \alpha\right\}. \tag{4}$$

Under conversation-level exchangeability, the resulting rule controls the marginal probability of an unwarned first failure at level α; Section 4 establishes the finite-sample guarantee and corresponding efficiency analysis. Full calibration details are provided in Appendix A.3.

4 Theoretical Results

We first establish the finite-sample validity of the calibrated warning threshold, and the proof is given in Appendix B.1.

Proposition 4.1 (Finite-grid missed-warning control)**.** Let Γ be a finite threshold grid with $\eta_1 < 0$, and suppose $n_{\text{cal}} \ge \lceil \alpha^{-1} \rceil - 1$. Conditional on the fitted score-construction pipeline, assume that the calibration and test loss functions $\{L_i(\cdot)\}_{i\in\mathcal{I}_{\text{cal}}}$ and $L_{\text{new}}(\cdot)$ are exchangeable. Then $\widehat{\eta}$ in equation 4 is well defined and $\mathbb{E}\left[L_{\text{new}}(\widehat{\eta})\right] = \mathbb{P}\left\{L_{\text{new}}(\widehat{\eta}) = 1\right\} \le \alpha$.

Missed-warning control alone is insufficient, since an always-warn rule trivially satisfies the constraint. For an arbitrary pre-response score $S(i,t)$, we define $V_{S,i} = \max_{1\le t\le Y_i} S(i,t)$, $W_S(\eta) = \mathbb{P}\{V_{S,i} > \eta\}$, and $R_S(\eta) = \mathbb{P}\{\Delta_i = 1, V_{S,i} \le \eta\}$. We study the constrained population objective

$$\min_{S,\eta} W_S(\eta) \qquad \text{subject to} \qquad R_S(\eta) \le \alpha.$$

Let $X_{i,t} = (t, \widehat{U}_{i,t}, B_{i,t})$, define the propagation-oblivious and propagation-aware score classes $\mathcal{S}_0 = \{S : S(i,t) = s(X_{i,t})\}$ and $\mathcal{S}_H = \{S : S(i,t) = s(X_{i,t}, H_{i,t})\}$. For a score class $\mathcal{S}$, let $W^*_{\mathcal{S}}(\alpha) = \inf_{S\in\mathcal{S},\, \eta: R_S(\eta)\le\alpha} W_S(\eta)$. In particular, $W^*_0(\alpha) = W^*_{\mathcal{S}_0}(\alpha)$ and $W^*_H(\alpha) = W^*_{\mathcal{S}_H}(\alpha)$. Since $\mathcal{S}_0 \subseteq \mathcal{S}_H$, $W^*_H(\alpha) \le W^*_0(\alpha)$, so access to propagated history cannot worsen the oracle warning-burden frontier. To connect this oracle comparison to the fitted rule, we state a finite-sample oracle inequality for the calibrated propagation-aware rule below. The full formal theorem, proof, and supporting technical lemmas are provided in Appendix B.

Theorem 4.2 (Informal oracle inequality)**.** Under the regularity conditions of Theorem B.1 and suppressing the deterministic threshold-grid discretization, the calibrated propagation-aware rule satisfies, with high probability,

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W_H^\star(\alpha) + b_H(\alpha) + O_{\mathbb{P}}\left(\sqrt{\frac{p_H}{n_{\mathrm{tr}}}} + \sqrt{\frac{\log|\Gamma|}{n_{\mathrm{cal}}}} + \frac{1}{n_{\mathrm{cal}}}\right),$$

where $b_H(\alpha)$ is the approximation gap of the population hazard-fitting target and p_H is the dimension of the propagation-aware hazard model.

Let $r_n^\star = \sqrt{p_H/n_{\mathrm{tr}}} + \sqrt{\log|\Gamma|/n_{\mathrm{cal}}} + n_{\mathrm{cal}}^{-1}$ collect the training-estimation and calibration errors, where $n_{\mathrm{tr}} = |\mathcal{I}_{\mathrm{tr}}|$ and $C < \infty$ is a constant depending on population quantities.

Definition 1 (Cascade-separation condition)**.** Fix a missed-warning level $\alpha \in (0, \pi)$ and a slack $\varepsilon_0 \ge 0$ such that $\alpha + \varepsilon_0 < \pi$. We say the warning problem exhibits *cascade separation* at level (α, ε_0) with margin $c_{\mathrm{sep}} > 0$ if the oracle warning burdens of the propagation-oblivious and propagation-aware score classes satisfy $W_0^*(\alpha + \varepsilon_0) - W_H^*(\alpha) \ge c_{\mathrm{sep}}$.

Definition 1 is an oracle-level margin condition whose role is to quantify the population-level burden separation between the two score classes, and Appendix A.4 provides a more interpretable sufficient condition. The oracle separation gap admits the decomposition

$$W_0^*(\alpha + \varepsilon_0) - W_H^*(\alpha) = \{W_0^*(\alpha) - W_H^*(\alpha)\} - \{W_0^*(\alpha) - W_0^*(\alpha + \varepsilon_0)\}.$$

Since $W_0^*(\cdot)$ is nonincreasing, the second term is nonnegative. Thus, ε_0 reduces the same-risk oracle gap by exactly the burden reduction afforded to the propagation-oblivious oracle by relaxing the missed-warning constraint from α to $\alpha + \varepsilon_0$. The margin c_{sep} therefore quantifies the remaining oracle advantage available to offset the approximation gap and the finite-sample error $Cr_n^\star$.

Corollary 1 (Burden reduction under cascade separation)**.** Suppose the conditions of Theorem B.1 hold, and that the warning problem satisfies the cascade-separation condition in Definition 1 at level (α, ε_0) with margin $c_{\mathrm{sep}} > 0$. Then, with high probability,

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W_0^*(\alpha + \varepsilon_0) - \{c_{\mathrm{sep}} - b_H(\alpha) - Cr_n^\star\}.$$

Hence, if $c_{\mathrm{sep}} > b_H(\alpha) + Cr_n^\star$, then the fitted propagation-aware warning rule achieves strictly smaller warning burden than the propagation-oblivious oracle operating at the relaxed missed-warning level $\alpha + \epsilon_0$. Figure 2 provides an empirical illustration of Corollary 1 by comparing two fitted scores obtained from the same pipeline. To quantify this separation empirically, we define $\hat{\epsilon}_{\max}$ as the largest slack such that the score with H, calibrated at α, retains no larger warning rate than the score without H calibrated at $\alpha + \epsilon$. A larger $\hat{\epsilon}_{\max}$ indicates a greater warning-burden advantage from incorporating propagated history.

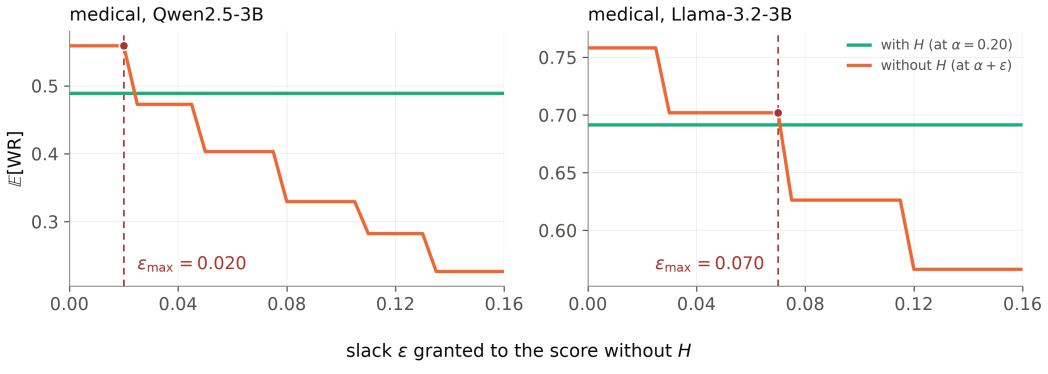

Figure 2: Empirical counterpart of cascade separation (Medical, $t \ge 3$), averaged over 20 splits.

5 Experimental Results

Data Generation and Setting. For multi-turn generation, we use 950 seed questions drawn from the four HalluHard domains (Fan et al., 2026). Each seed question initiates a six-turn

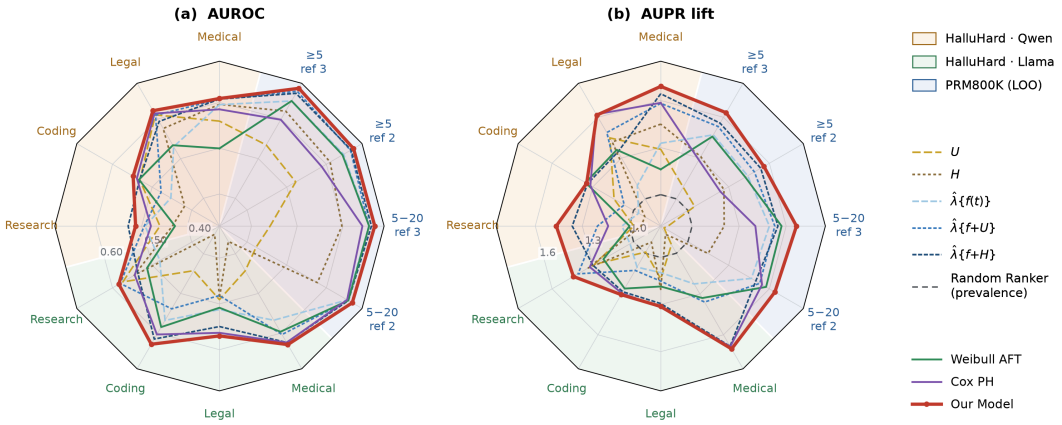

Figure 3: Sequential first-failure prediction on conversations that survive to turn 3 and on the PRM800K LOO settings. (a) AUROC. (b) AUPR lift (i.e., AUPR/π with π the prevalence). Entries are means over 20 random splits, conversation-level for HalluHard and problem-level for PRM800K.

conversation, with five follow-up queries generated under a domain-specific role prompt. We use `Qwen2.5-3B-Instruct` and `Llama-3.2-3B-Instruct` as target models, yielding eight domain model cells. First factual failures are identified by larger same-family judges (`Qwen2.5-7B-Instruct` and `Llama-3.1-8B-Instruct`): $E_{i,t} = 1$ if the response contains at least one claim labeled `incorrect`, while `uncertain` claims are treated as failure-free. At each at-risk state, we draw $K = 5$ diagnostic responses to construct the uncertainty features. For multi-step reasoning, we use PRM800K (Lightman et al., 2024), which contains approximately 800K human step-level correctness labels across 75K model-generated solutions to 12K math problems. We define the first step rated -1 as the first reasoning failure. Uncertainty is estimated from leave-one-out sibling chains that remain active at the current step, requiring no additional generation. To scale semantic grouping, we use exact match after text normalization in place of pairwise entailment clustering. Unless otherwise noted, the theoretical and conformal analyses use the capped regime $5 \le K_p \le 20$. We split conversations (problems for PRM800K) into training, calibration, and test sets in a $60/20/20$ ratio and repeat the full fitting and calibration pipeline over 20 random splits. Further details on generation, judging, semantic grouping, labeling stability, and robustness to the PRM800K reference-pool construction are provided in Appendix C.1. Because H is designed to capture accumulated pre-failure history, our primary HalluHard analysis focuses on conversations remaining at risk at turn 3, with evaluation over subsequent at-risk turns. The full *numerical results* underlying all figures, including the corresponding full-risk-set results, are deferred to Appendix C.

Evaluation Metrics and Comparison Models. We evaluate both first-failure ranking and decision-level warning efficiency. For ranking, we report **AUROC** and **AUPR** over the at-risk set using $Z_{i,t} = \mathbf{1}\{Y_i = t, \Delta_i = 1\}$. For decision making, we consider the missed-warning rate (MWR), defined empirically as $\mathrm{MWR} = \frac{1}{n_{\mathrm{test}}} \sum_{i \in I_{\mathrm{test}}} \mathbf{1}\{\Delta_i = 1, \tau_i^{\mathrm{warn}} > \tau_i\}$, which is the fraction of test units for which an observed first failure occurs without a warning at or before the failure. We treat MWR as a hard constraint at level α. Among admissible constructions, warning burden is measured by the **warning rate** (WR) and **unnecessary-warning rate** (UWR), with lower values preferred. Formal definitions are given in Appendix C.1. We first compare our method with the semantic entropy U (Kuhn et al., 2023), the propagated instability feature H, and their rank-normalized average. We then compare against restricted hazard models using these features, namely $\widehat{\lambda}\{f(t)\}$, $\widehat{\lambda}\{f+U\}$, and $\widehat{\lambda}\{f+H\}$. Finally, we hold the feature set fixed and vary the survival model, using Weibull AFT and Cox PH. Additional comparisons are reported in Appendix C. All constructions use the same splits, diagnostic samples, threshold grid, and conformal calibration procedure.

Experimental Results. We then present our experimental results, which are aimed at answering the following questions. **Q1.** Does the propagation-aware hazard score rank first failures better than existing uncertainty scores and restricted feature sets? **Q2.** Does the hazard layer contribute beyond the features it aggregates? **Q3.** Under a common missed-warning target, does our full warning rule attain lower warning burden than alternative score constructions and survival-model baselines? In

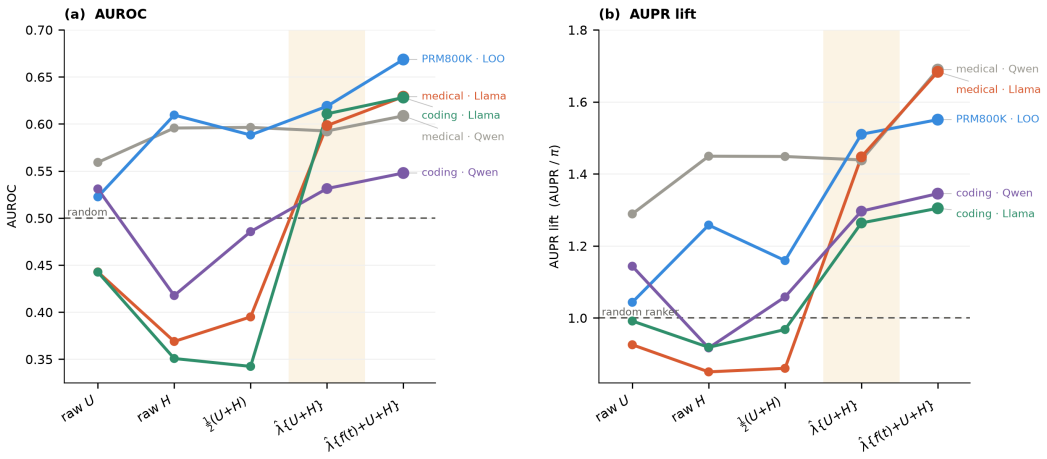

Figure 4: An ablation of score constructions for turn-level failure prediction.

addition to these main questions, we also evaluate the effect of H across different hazard modeling approaches in Appendix C.2.

To answer Q1, Figure 3 reports AUROC and AUPR lift (AUPR divided by the fold prevalence) across twelve settings. We first vary the feature set under a fixed hazard model. Our full construction $\widehat{\lambda}\{f(t)+U+H\}$ attains the highest point-estimate AUPR lift in all twelve settings, and the highest AUROC in three of the four Qwen settings and in all four Llama and all four PRM800K settings. However, in Research (Qwen), the construction without U, $\widehat{\lambda}\{f(t)+H\}$, achieves a slightly higher AUROC, whereas our full construction achieves the higher AUPR. This pattern can be explained in part by the current-turn uncertainty feature being both saturated and marginally uninformative. With only five diagnostic samples, $\widehat{U}$ takes only a small number of discrete values. As shown in Table A1, nearly half of the at-risk states in Research (Qwen) attain the maximum value of $\widehat{U}$, and the mean U is nearly identical between first-failure and non-failure turns. We can also see that Coding (Qwen) is comparably saturated but retains a clear mean gap. Thus, U provides little marginal discrimination in Research (Qwen), which helps explain why adding it to $f(t)+H$ does not improve AUROC. We then hold the feature set fixed at $f(t)+U+H$ and vary the survival model. Our discrete-time hazard model attains the highest AUROC in all settings and the highest AUPR lift in 9, while remaining comparable to the best-performing method in the two legal settings and coding setting (Qwen).

Having established the consistent incremental gain from H within our discrete-time hazard model, we next examine how the hazard layer combines current and historical uncertainty signals. We perform an ablation study on two contrasting HalluHard domains, Medical and Coding, together with a PRM800K LOO setting. Figure 4 decomposes the progression from the raw current and historical uncertainty summaries to the learned first-failure score. We first observe that the raw instability summaries contain predictive structure, although their association with first-failure risk is not uniformly positive. Across the examples shown, the historical feature provides complementary predictive information beyond the current-turn uncertainty, but this information is not necessarily aligned in the same direction. Consequently, simply combining U and H with fixed weights does not consistently resolve this heterogeneity. In contrast, passing U and H through a learned hazard layer, even without the baseline term $f(t)$, allows their directions and relative contributions to be learned from data and produces a score that is more consistently aligned with first-failure risk. Adding the baseline turn effect then yields the full score and provides a further improvement in most settings. Thus, we find that the gain comes from learning how current and propagated instability jointly relate to first-failure risk, together with the baseline temporal structure of the sequential process.

Overall, these results separate two findings. Within our discrete-time hazard model, adding H improves both AUROC and AUPR in all twelve evaluated settings. Across alternative survival-model families, however, its contribution is more heterogeneous and depends on the domain and model family. Thus, the predictive value of propagated historical instability is not specific to our hazard specification, although the magnitude of its benefit is model and setting dependent.

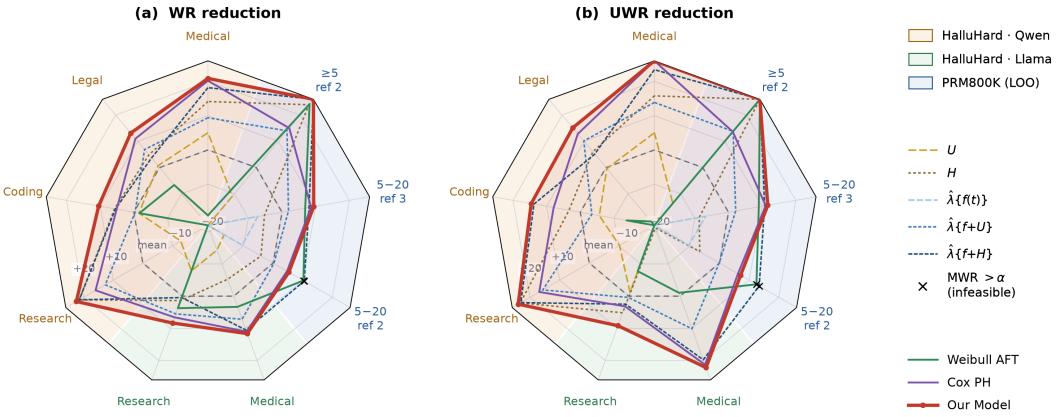

Figure 5: Conformal results at $\alpha = 0.20$ for HalluHard landmarks and PRM800K LOO. (a) WR reduction; (b) UWR reduction. Values are percentage reductions relative to the mean across constructions; larger is better. Crosses mark infeasible methods. Means over 20 random splits.

We next turn from predictive performance to Q3. Due to space limitations, the main text reports only the common missed-warning target $\alpha = 0.20$, in terms of the reduction in warning rate (WR) and unnecessary-warning rate (UWR) (Figure 5). Full numerical results, including $\alpha = 0.10$ and $\alpha = 0.05$, are in Appendix C. We can observe that our full model traces the outermost feasible contour in both panels, attaining the lowest WR and the lowest UWR among all feasible constructions across the six HalluHard landmark cells satisfying the calibration-size requirement and the three PRM800K LOO settings. The improvement is particularly visible in Coding and Research, while Medical and Legal exhibit more moderate gains. The reduction is more pronounced in UWR, indicating that the burden reduction is especially strong among conversations that remain failure-free. However, in some PRM800K settings, Weibull AFT with H and $\widehat{\lambda}\{f(t) + H\}$ can attain slightly lower WR and UWR than our full model while realizing infeasible MWR. This pattern is expected because their lower warning burden is obtained by missing more first failures. These infeasible operating points therefore do not establish an advantage at matched missed-warning risk. Concretely, on PRM800K ($5 \leq K_p \leq 20$, min-ref of 2) at $\alpha = 0.20$, $\widehat{\lambda}\{f(t) + H\}$ and Weibull AFT with H realize $\mathrm{MWR} = 0.2036$ and 0.2029, against 0.1994 for our full model, so their overshoots relative to our realized risk are 0.0042 and 0.0035. Since WR is nonincreasing and MWR nondecreasing in the warning threshold (Lemma 4), tolerating additional missed first failures necessarily lowers the warning burden, and the exchange rate along our own calibrated path is $\Delta\mathrm{WR}/\Delta\mathrm{MWR} = -3.1$ over $\alpha \in [0.10, 0.20]$. Because $|d\mathrm{WR}/d\mathrm{MWR}|$ is decreasing in the threshold (Appendix A.5), this chord upper-bounds the local rate at $\alpha = 0.20$. The two MWR overshoots can therefore induce WR reductions of up to 0.0130 and 0.0109, respectively, both at least as large as the observed differences of 0.0109 and 0.0103. Thus, the apparent WR advantages are within what can be explained by the mismatch in realized MWR, which these two constructions fail to control at the target level. The same conclusion holds for UWR.

6 Discussion

In this paper, we formulate prospective first-failure monitoring as a distinct problem from repeated response-level verification. We further introduce a historical instability feature that propagates pre-failure information, since within the at-risk process historical information cannot take the form of past observed failures. Beyond predictive modeling, we provide finite-sample missed-warning control through conformal calibration and establish an oracle inequality that characterizes when the additional historical information can translate into lower warning burden after accounting for approximation and estimation errors. Empirically, the propagated feature improves both AUROC and AUPR in all twelve evaluated settings within our discrete-time hazard model, while its benefit across alternative survival-model families is more heterogeneous. Together, these results suggest that modeling accumulated pre-failure instability can provide useful complementary information for prospective first-failure monitoring.

AI USE STATEMENT:

We used an LLM (ChatGPT) for proofreading paper, for checking our mathematical derivations, and as part of the data generation for the multi-turn conversations. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

REFERENCES

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Chenkang, Fan Yu, Junjie Nian, Sihan Zhao, Zhuoka Feng, Zijun Yao, Wang Heng, Yu Minshen, and Yixin Cao. Thinking traps in long chain-of-thought: A measurable study and trap-aware adaptive restart. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38746–38760, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1930. URL https://aclanthology.org/2026.findings-acl.1930/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960. URL https://api.semanticscholar.org/CorpusID:15926286.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x.

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. Halluhard: A hard multi-turn hallucination benchmark, 2026. URL https://arxiv.org/abs/2602.01031.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Ashley Lewis, Andrew Perrault, Eric Fosler-Lussier, and Michael White. VISTA: Verification in sequential turn-based assessment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40685–40714, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1890. URL https://aclanthology.org/2026.acl-long.1890/.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6679–6700, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.347. URL https://aclanthology.org/2025.findings-acl.347/.

Yubo Li, Ramayya Krishnan, and Rema Padman. Time-to-inconsistency: A survival analysis of large language model robustness to adversarial attacks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=88C7vSdn0t.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RwzFNbJ3Ez.

G Miller. Note on the bias of information estimates. 1955. URL https://api.semanticscholar.org/CorpusID:125662170.

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. Hard2Verify: A step-level verification benchmark for open-ended frontier math. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22502–22517, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1031. URL https://aclanthology.org/2026.acl-long.1031/.

Judith D. Singer and John B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195, 1993. ISSN 03629791. URL http://www.jstor.org/stable/1165085.

Zvi Topol. Quantifying LLM safety degradation under repeated attacks using survival analysis. In Kai-Wei Chang, Ninareh Mehrabi, Satyapriya Krishna, Anubrata Das, Jwala Dhamala, Yang Trista Cao, Tharindu Kumarage, Anil Ramakrishna, Christos Christodoulopoulos, Yixin Wan, Aram Galystan, Anoop Kumar, and Rahul Gupta (eds.), *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 64–72, San Diego, California, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-418-7. doi: 10.18653/v1/2026.trustnlp-main.5. URL https://aclanthology.org/2026.trustnlp-main.5/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang, and Nigel Collier. Confidence estimation for LLMs in multi-turn interactions. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25661–25676, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1280. URL https://aclanthology.org/2026.findings-acl.1280/.

Zhaohan Zhang, Chengzhengxu Li, Xiaoming Liu, Chao Shen, Ziquan Liu, and Ioannis Patras. Confidence should be calibrated more than one turn deep. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38578–38594, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1787. URL https://aclanthology.org/2026.acl-long.1787/.

APPENDIX

A TECHNICAL DETAILS

A.1 CHAIN-OF-THOUGHT EXTENSION

Problem Setting: Let problem p have prompt x_p and K_p reasoning chains independently sampled from the LLM conditional on x_p. Chain k of problem p is represented by $\mathcal{T}_{p,k} = \big(a_{p,k,1}, a_{p,k,2}, \ldots, a_{p,k,T_{p,k}}\big)$, where $a_{p,k,t}$ is the reasoning step produced at position t and $T_{p,k}$ is the observed chain length. Define the step-level error indicator $E_{p,k,t} = \mathbf{1}\{a_{p,k,t} \text{ contains at least one reasoning error}\}$, and the first reasoning failure time $\tau_{p,k} = \inf\{t : E_{p,k,t} = 1\}$, with the convention $\inf \varnothing = \infty$. Unlike the multi-turn setting, where uncertainty is estimated by resampling from the conversation's own pre-response history, the chain-of-thought monitor uses sibling chains for the same problem as an offline reference pool. For prediction at step t, define the target chain's pre-step history $\mathcal{F}_{p,k,t} = (x_p, a_{p,k,1}, \ldots, a_{p,k,t-1})$, and the leave-one-chain-out sibling index set $\mathcal{J}^{(-k)}_{p,k,t} = \{k' \neq k : T_{p,k'} \geq t\}$, with the corresponding reference pool $\mathcal{A}^{(-k)}_{p,k,t} = (a_{p,k',t})_{k' \in \mathcal{J}^{(-k)}_{p,k,t}}$. We then define the augmented pre-step information set $\mathcal{G}_{p,k,t} = \sigma(\mathcal{F}_{p,k,t}, \mathcal{J}^{(-k)}_{p,k,t}, \mathcal{A}^{(-k)}_{p,k,t})$. Thus, $\mathcal{G}_{p,k,t}$ contains the target chain's history through step $t-1$ together with the available sibling outputs at position t, but does not contain the target output $a_{p,k,t}$. All monitoring quantities in the chain-of-thought setting are therefore defined with respect to $\mathcal{G}_{p,k,t}$.

Because the sibling reference set must contain a minimum number of chains, monitoring may become unavailable before the target reasoning chain itself ends. Let $n_{p,k,t} = |\mathcal{J}^{(-k)}_{p,k,t}|$ denote the reference-pool size, and let $n_{\min}$ be the minimum pool size required by the monitoring procedure in later section. Since sibling chains may terminate but do not resume, $n_{p,k,t}$ is nonincreasing in t. Monitorability is therefore prefix-closed. We define the observed monitoring horizon by $C_{p,k} = \max\{t \leq T_{p,k} : n_{p,k,t} \geq n_{\min}\}$, with the convention $C_{p,k} = 0$ if the set is empty. Hence, $C_{p,k} \leq T_{p,k}$ is the last step at which the information required by the monitoring procedure is available. We observe the monitored failure-time pair $Y_{p,k} = \min(\tau_{p,k}, C_{p,k})$ and $\Delta_{p,k} = \mathbf{1}\{\tau_{p,k} \leq C_{p,k}\}$. Thus, even among reasoning chains that reach step t, only those satisfying $C_{p,k} \geq t$ remain within the observable monitoring process. Accordingly, our target in the chain-of-thought setting is the first-failure hazard on the monitorable process,

$$\lambda_{p,k,t} = \mathbb{P}\left(\tau_{p,k} = t \mid \tau_{p,k} \geq t, C_{p,k} \geq t, \mathcal{G}_{p,k,t}\right). \tag{5}$$

Because $C_{p,k} \geq t$ is equivalent, for a chain reaching step t, to the observed condition $n_{p,k,t} \geq n_{\min}$, and $n_{p,k,t}$ is $\mathcal{G}_{p,k,t}$-measurable, monitorability is part of the observed pre-step state. We therefore target first-failure risk over the observable monitorable process and do not extrapolate to steps for which the sibling reference pool is insufficient. The corresponding at-risk set is $\mathcal{R} = \{(p,k,t) : 1 \leq t \leq Y_{p,k}\}$, and, for each $(p,k,t) \in \mathcal{R}$, the event label is $Z_{p,k,t} = \mathbf{1}\{Y_{p,k} = t, \Delta_{p,k} = 1\}$. Since membership in $\mathcal{R}$ implies $\tau_{p,k} \geq t$, we have $E_{p,k,s} = 0$ for all $s < t$. Thus, as in the multi-turn setting, observed past failures cannot serve as historical predictors within the risk set. Historical information must instead be propagated through a pre-failure instability signal, which motivates the semantic-dispersion and propagation features introduced in the following section.

Method: For each chain-step triple (p,k,t) with $t \leq T_{p,k}$, define the leave-one-out sibling index set $\mathcal{J}^{(-k)}_{p,k,t} = \{k' \neq k : T_{p,k'} \geq t\}$, and the corresponding **leave-one-out reference pool**

$$\mathcal{A}^{(-k)}_{p,k,t} = \big(a_{p,k',t}\big)_{k' \in \mathcal{J}^{(-k)}_{p,k,t}},$$

collecting the step-t outputs of all sibling chains still active at step t, with pool size $n_{p,k,t} = \left|\mathcal{J}^{(-k)}_{p,k,t}\right|$. We require $n_{p,k,t} \geq n_{\min}$ for meaningful uncertainty estimation and define the monitoring horizon

$$C_{p,k} = \max\big\{t \leq T_{p,k} : n_{p,k,t} \geq n_{\min}\big\},$$

for chains for which this set is nonempty. Since $n_{p,k,t}$ is nonincreasing in t, the monitorable steps are prefix-closed. The observed pair $(Y_{p,k}, \Delta_{p,k})$ and the at-risk set $\mathcal{R}$ are then defined as.

The sibling chains in $\mathcal{J}_{p,k,t}^{(-k)}$ are grouped into semantic equivalence classes $\mathcal{C}_{p,k,t,1}, \ldots, \mathcal{C}_{p,k,t,m_{p,k,t}}$, where each class contains the indices of sibling chains whose step-t outputs are semantically equivalent. The corresponding empirical proportions are $\widehat{p}_{p,k,t,c} = \frac{|\mathcal{C}_{p,k,t,c}|}{n_{p,k,t}}$. The semantic dispersion score $\widehat{U}_{p,k,t}$ and non-modal mass $\widehat{q}_{p,k,t}$ are computed using the same functional forms as in Section 3, with $n_{p,k,t}$ replacing K. The Miller–Madow correction uses $n_{p,k,t}$ as the sample size. The propagation score $H_{p,k,t}(\widehat{\rho})$ and hazard model $\lambda_\theta(p,k,t)$ proceed analogously with (i,t) replaced by (p,k,t) throughout.

For conformal calibration, the training, calibration, and test splits are performed at the problem level, with all sibling chains from the same problem assigned to the same split. For each monitored chain (p,k), define the chain-level missed-warning loss analogously to the multi-turn setting,

$$L_{p,k}(\eta) = \Delta_{p,k}\mathbf{1}\left\{\max_{1\le t\le Y_{p,k}} R_{p,k,t} \le \eta\right\}.$$

Let $\mathcal{K}_p$ denote the set of monitored chains for problem p. We aggregate the chain-level losses within each problem as

$$L_p(\eta) = \max_{k\in\mathcal{K}_p} L_{p,k}(\eta).$$

Thus, $L_p(\eta) = 1$ if at least one monitored chain for problem p experiences an observed first reasoning failure without a warning before or at that failure. Conformal calibration is then performed using the problem-level losses $\{L_p(\eta)\}$, so that problems form the exchangeable calibration units. Accordingly, Proposition 4.1 applies with the conversation-level loss $L_i(\eta)$ replaced by the problem-level loss $L_p(\eta)$. Unlike live resampling in the multi-turn setting, this leave-one-out construction uses already-generated sibling chains and therefore incurs no additional LLM generation cost; its monitoring horizon is instead limited by the shrinking sibling pool.

A.2 Semantic Uncertainty Construction

For each at-risk pair $(i,t) \in \mathcal{R}$, we draw K responses, that is $\tilde{a}_{i,t}^{(1)}, \ldots, \tilde{a}_{i,t}^{(K)} \sim p_{\mathrm{LLM}}(\cdot \mid \mathcal{F}_{i,t})$. The past responses in $\mathcal{F}_{i,t}$ are the actual previous outputs and are not resampled. The K diagnostic responses are grouped into semantic equivalence classes $C_{i,t,1}, C_{i,t,2}, \ldots, C_{i,t,m_{i,t}}$. Using all K responses, let $\widehat{p}_{i,t,c} = K^{-1}|C_{i,t,c}|$, where $c = 1, \ldots, m_{i,t}$. We define the current semantic dispersion score using the Miller-Madow-corrected entropy estimator (Miller, 1955), that is

$$\widehat{U}_{i,t} = -\sum_{c=1}^{m_{i,t}} \widehat{p}_{i,t,c} \log \widehat{p}_{i,t,c} + \frac{m_{i,t}-1}{2K}. \tag{6}$$

Then, we define the non-modal semantic mass

$$\widehat{q}_{i,t} = 1 - \max_{1\le c\le m_{i,t}} \widehat{p}_{i,t,c}. \tag{7}$$

Thus, $\widehat{U}_{i,t}$ measures current semantic dispersion, whereas $\widehat{q}_{i,t} \in [0,1]$ measures the lack of concentration around the modal semantic answer. We use $\widehat{U}_{i,t}$ as a current-turn feature and propagate $\widehat{q}_{i,t}$ through the history. Because K is fixed across states, all uncertainty features are computed under the same diagnostic sampling budget.

A.3 Details for Calibrated Warning Thresholds

The same construction applies to the chain-of-thought setting, with chains replacing conversations as survival units and leave-one-out sibling chains replacing live resampling as the uncertainty source. Details are given in Appendix A.1.

Given the estimated hazard score, we can rank pre-response states by failure risk. To translate these scores into warnings, we introduce a threshold η, and trigger a warning whenever $R_{i,t} > \eta$. The key remaining question is how to choose such a threshold in a statistically principled way. In this paper, we adopt conformal risk control to calibrate this threshold. Because $R_{i,t}$ is computed before releasing response $a_{i,t}$, a warning issued at the failure turn $t = \tau_i$ is considered successful. Hence, for a fixed threshold η, the conversation-level missed-warning loss is

$$L_i(\eta) = \Delta_i \mathbf{1}\Big\{\max_{1\le t\le Y_i} R_{i,t} \le \eta\Big\}.$$

Since increasing η makes warnings less frequent, $L_i(\eta)$ is nondecreasing in η. We split the conversations into three disjoint sets, $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{cal}}, \mathcal{I}_{\text{test}}$. The hazard model and the entire score-construction pipeline are trained only on $\mathcal{I}_{\text{train}}$. After training, the score-construction procedure is fixed before calibration. Let $\Gamma = \{\eta_1 < \eta_2 < \cdots < \eta_m\}$ be a finite grid of candidate thresholds. Because $R_{i,t} \in (0,1)$, we include a conservative endpoint $\eta_1 < 0$ in Γ for which $L_i(\eta_1) = 0$ for every calibration conversation. Define the empirical calibration risk $\widehat{R}_{\text{cal}}(\eta) = n_{\text{cal}}^{-1} \sum_{i \in \mathcal{I}_{\text{cal}}} L_i(\eta)$, and we assume $n_{\text{cal}} \geq \lceil 1/\alpha \rceil - 1$, where $n_{\text{cal}} = |\mathcal{I}_{\text{cal}}|$. η_1 always satisfies the corrected calibration criterion below, ensuring that the feasible set is nonempty. We choose the largest threshold whose finite-sample corrected calibration risk is below the target level α:

$$\widehat{\eta} = \max \left\{ \eta \in \Gamma : \frac{n_{\text{cal}}}{n_{\text{cal}} + 1} \widehat{R}_{\text{cal}}(\eta) + \frac{1}{n_{\text{cal}} + 1} \leq \alpha \right\}. \tag{8}$$

By monotonicity, $\widehat{\eta}$ is the largest, and hence least conservative, threshold satisfying the corrected calibration bound. Under conversation-level exchangeability, the calibrated rule controls the marginal missed-warning risk at level α; we establish this guarantee, together with the corresponding efficiency analysis, in Section 4.

The method developed above applies to the chain-of-thought setting. The survival unit changes from conversation i to chain (p, k), and the uncertainty source changes from live resampling to leave-one-out cross-chain comparison. Due to space limitation, we defer the details to Appendix A.1.

A.4 A Gaussian Surrogate for the Separation Margin

For the Gaussian surrogate, we reduce each variable-length at-risk feature trajectory to a fixed-dimensional conversation-level summary. Let $G_{0,i} \in \mathbb{R}^p$ be a conversation-level summary built from the propagation-oblivious trajectory $\{X_{i,t} : 1 \leq t \leq Y_i\}$, and let $H_i^\star \in \mathbb{R}$ be a scalar conversation-level propagated-instability summary. We use an additive aggregate, for example the at-risk average $H_i^\star = \frac{1}{Y_i} \sum_{t=1}^{Y_i} H_{i,t}$, chosen so that a Gaussian surrogate is plausible; the terminal value H_{i,Y_i} is an equally admissible choice. Define the propagation-aware summary $G_{H,i} = (G_{0,i}^\top, H_i^\star)^\top \in \mathbb{R}^{p+1}$. Let $D_i = \Delta_i$ and $\pi = \mathbb{P}(D_i = 1) \in (0,1)$. Conditional on $D_i = d$, $d \in \{0,1\}$, we assume

$$G_{H,i} \mid D_i = d \sim N(\mu_d, \Sigma), \qquad \Sigma \succ 0,$$

where $\mu_d = \begin{pmatrix} \mu_{0d} \\ \mu_{Hd} \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{0H} \\ \Sigma_{H0} & \Sigma_{HH} \end{pmatrix}$, with $G_{0,i} \in \mathbb{R}^p$ and $H_i^\star \in \mathbb{R}$, and set $\delta_0 = \mu_{01} - \mu_{00}$ and $\delta_H = \mu_{H1} - \mu_{H0}$.

Under this surrogate, the propagation-oblivious oracle frontier is represented by the Neyman–Pearson discriminant based on $G_{0,i}$, whereas the propagation-aware oracle frontier is represented by the Neyman–Pearson discriminant based on $G_{H,i}$. We want to note that this surrogate is used only for the Gaussian burden calculation and is not required for our previous results (i.e., conformal validity and the nesting-based oracle dominance). For a Gaussian summary with class-mean difference δ and common covariance Σ, the Mahalanobis separation is $D^2 = \delta^\top \Sigma^{-1} \delta$. For the two summaries,

$$D_0^2 = \delta_0^\top \Sigma_{00}^{-1} \delta_0, \qquad D_H^2 = (\mu_1 - \mu_0)^\top \Sigma^{-1} (\mu_1 - \mu_0).$$

Define the residual historical-instability shift and residual variance

$$\delta_{H|0} = \delta_H - \Sigma_{H0} \Sigma_{00}^{-1} \delta_0, \qquad \sigma_{H|0}^2 = \Sigma_{HH} - \Sigma_{H0} \Sigma_{00}^{-1} \Sigma_{0H} > 0.$$

By the Schur-complement identity, $D_H^2 = D_0^2 + \frac{\delta_{H|0}^2}{\sigma_{H|0}^2}$. Hence $D_H \geq D_0$, with strict inequality if and only if $\delta_{H|0} \neq 0$. In words, $H_i^\star$ increases the discriminant separation precisely when it carries a residual class shift after linearly adjusting for the propagation-oblivious summary $G_{0,i}$. For $0 < a < \pi$, define the Gaussian oracle burden function

$$\mathcal{W}(D, a) = (\pi - a) + (1 - \pi) \Phi\left(-D - \Phi^{-1}\left(\tfrac{a}{\pi}\right)\right),$$

where Φ is the standard normal distribution function. Under the surrogate, $\mathcal{W}(D, a)$ is the oracle warning burden at missed-warning level a for a summary with Mahalanobis separation D. Since the term $\pi - a$ does not depend on D and $D \mapsto \Phi(-D - \Phi^{-1}(a/\pi))$ is strictly decreasing, $\mathcal{W}(D, a)$ is strictly decreasing in D for each fixed $a \in (0, \pi)$.

Proposition A.1 (Verifiable Gaussian criterion for cascade separation)**.** Under the Gaussian discriminant surrogate described above, fix $\alpha \in (0, \pi)$ and $\varepsilon_0 \geq 0$ with $\alpha + \varepsilon_0 < \pi$. Under the Gaussian surrogate, the cascade-separation gap satisfies

$$W_0^*(\alpha + \varepsilon_0) - W_H^*(\alpha) = c_G(\alpha, \varepsilon_0),$$

where $c_G(\alpha, \varepsilon_0)$ is given by

$$\begin{aligned} c_G(\alpha, \varepsilon_0) &= \mathcal{W}(D_0, \alpha + \varepsilon_0) - \mathcal{W}(D_H, \alpha) \\ &= -\varepsilon_0 + (1 - \pi)\left[\Phi\left(-D_0 - \Phi^{-1}\left(\frac{\alpha + \varepsilon_0}{\pi}\right)\right) - \Phi\left(-D_H - \Phi^{-1}\left(\frac{\alpha}{\pi}\right)\right)\right]. \end{aligned}$$

Therefore, the cascade-separation condition in Definition 1 holds with any margin $0 < c_{\text{sep}} \leq c_G(\alpha, \varepsilon_0)$ whenever $c_G(\alpha, \varepsilon_0) > 0$.

Proof. Under the Gaussian surrogate and the Gaussian burden formula,

$$W_0^*(\alpha + \varepsilon_0) = \mathcal{W}(D_0, \alpha + \varepsilon_0), \qquad W_H^*(\alpha) = \mathcal{W}(D_H, \alpha).$$

Therefore, $W_0^*(\alpha + \varepsilon_0) - W_H^*(\alpha) = \mathcal{W}(D_0, \alpha + \varepsilon_0) - \mathcal{W}(D_H, \alpha) = c_G(\alpha, \varepsilon_0)$. If $c_G(\alpha, \varepsilon_0) > 0$, then for any $0 < c_{\text{sep}} \leq c_G(\alpha, \varepsilon_0)$,

$$W_0^*(\alpha + \varepsilon_0) - W_H^*(\alpha) \geq c_{\text{sep}}.$$

This is exactly the cascade-separation condition in Definition 1. Expanding the definition of $\mathcal{W}$ gives the stated closed-form expression for $c_G(\alpha, \varepsilon_0)$. □

Remark 1 (No-slack special case and testability)**.** If $\varepsilon_0 = 0$, then

$$c_G(\alpha, 0) = (1 - \pi)\left[\Phi\left(-D_0 - \Phi^{-1}\left(\tfrac{\alpha}{\pi}\right)\right) - \Phi\left(-D_H - \Phi^{-1}\left(\tfrac{\alpha}{\pi}\right)\right)\right].$$

Since $D \mapsto \Phi(-D - \Phi^{-1}(\alpha/\pi))$ is strictly decreasing, $c_G(\alpha, 0) > 0$ if and only if $D_H > D_0$, equivalently $\delta_{H|0} \neq 0$. Thus, without slack, any nonzero residual class shift in $H_i^\star$ suffices for strict oracle burden reduction. With $\varepsilon_0 > 0$ the condition is strictly stronger: the gain from H must overcome the relaxation ε_0 granted to the propagation-oblivious oracle. Finally, $\delta_{H|0}/\sigma_{H|0}^2$ is exactly the coefficient on $H_i^\star$ in the linear discriminant direction $\Sigma^{-1}(\mu_1 - \mu_0)$, equivalently the population coefficient on $H_i^\star$ in the logistic regression of Δ_i on $(G_{0,i}, H_i^\star)$; hence $\delta_{H|0} = 0$ is testable by a likelihood-ratio test.

A.5 Sensitivity Analysis

In this section, we study the relation between warning burden and missed-warning risk along the threshold path of a fixed score. Fix a score S and recall $V_{S,i} = \max_{1 \leq t \leq Y_i} S(i, t)$, with $W_S(\eta) = \mathbb{P}(V_{S,i} > \eta)$ and $R_S(\eta) = \mathbb{P}(\Delta_i = 1, V_{S,i} \leq \eta)$, and let $\pi = \mathbb{P}(\Delta_i = 1)$. We additionally recall the unnecessary-warning rate

$$\mathrm{UWR}_S(\eta) = \mathbb{P}\left(V_{S,i} > \eta \mid \Delta_i = 0\right).$$

Lemma 1 (Burden decomposition)**.** For every $\eta \in \mathbb{R}$, we have $W_S(\eta) = \pi - R_S(\eta) + (1 - \pi)\mathrm{UWR}_S(\eta)$.

Proof. Conditioning on Δ_i gives

$$W_S(\eta) = \pi \mathbb{P}(V_{S,i} > \eta \mid \Delta_i = 1) + (1 - \pi)\mathrm{UWR}_S(\eta).$$

Moreover, $R_S(\eta) = \pi \mathbb{P}(V_{S,i} \leq \eta \mid \Delta_i = 1) = \pi - \pi \mathbb{P}(V_{S,i} > \eta \mid \Delta_i = 1)$. Substituting the second identity into the first proves the claim. □

We want to note that the decomposition is purely algebraic and requires no distributional or shape assumption. In particular, for any $\eta_- < \eta_+$,

$$\begin{aligned} W_S(\eta_-) - W_S(\eta_+) = R_S(\eta_+) - R_S(\eta_-) \\ + (1 - \pi)\left\{\mathrm{UWR}_S(\eta_-) - \mathrm{UWR}_S(\eta_+)\right\}. \end{aligned}$$

Since $\mathrm{UWR}_S(\eta)$ is nonincreasing in η, this immediately gives $W_S(\eta_-) - W_S(\eta_+) \geq R_S(\eta_+) - R_S(\eta_-)$. Thus, without any monotone-likelihood-ratio assumption, increasing missed-warning risk by δ reduces warning burden by at least δ; the remaining reduction is exactly the weighted decrease in unnecessary warnings. To obtain an upper bound on finite burden changes, we impose the following smoothness and shape condition.

Assumption 1 (Smooth monotone likelihood ratio)**.** Assume $0 < \pi < 1$. On a threshold interval $\mathcal{I}$ of interest, the conditional laws of $V_{S,i}$ given $\Delta_i = d$, $d \in \{0, 1\}$, are absolutely continuous with densities f_d. Moreover, $f_1(\eta) > 0$ for every $\eta \in \mathcal{I}$, and $\rho(\eta) = \frac{f_0(\eta)}{f_1(\eta)}$ is nonincreasing in η on $\mathcal{I}$.

The absolute-continuity and positivity conditions ensure that $R_S(\eta) = \pi F_1(\eta)$ is strictly increasing on $\mathcal{I}$. Hence its image $\mathcal{A}_S = R_S(\mathcal{I})$ is an interval and R_S^{-1} is well defined on $\mathcal{A}_S$. We can therefore write warning burden and unnecessary-warning burden as functions of realized missed-warning risk,

$$\mathcal{W}_S(a) = W_S\left(R_S^{-1}(a)\right), \qquad \mathcal{U}_S(a) = \mathrm{UWR}_S\left(R_S^{-1}(a)\right), \qquad a \in \mathcal{A}_S.$$

Lemma 2 (Exchange-rate identities)**.** For every $a \in \mathcal{A}_S$, writing $\eta = R_S^{-1}(a)$, we have

$$\mathcal{W}_S'(a) = -\left(1 + \frac{(1-\pi)f_0(\eta)}{\pi f_1(\eta)}\right), \qquad \mathcal{U}_S'(a) = -\frac{f_0(\eta)}{\pi f_1(\eta)}.$$

Consequently,

$$|\mathcal{W}_S'(a)| = 1 + \frac{(1-\pi)f_0(\eta)}{\pi f_1(\eta)} \geq 1, \qquad |\mathcal{U}_S'(a)| = \frac{f_0(\eta)}{\pi f_1(\eta)},$$

and $|\mathcal{W}_S'(a)| = 1 + (1-\pi)|\mathcal{U}_S'(a)|$.

Proof. Since $R_S(\eta) = \pi F_1(\eta)$, we have $R_S'(\eta) = \pi f_1(\eta)$. Similarly, $\mathrm{UWR}_S(\eta) = 1 - F_0(\eta)$ gives $\mathrm{UWR}_S'(\eta) = -f_0(\eta)$, while

$$W_S(\eta) = 1 - \pi F_1(\eta) - (1-\pi)F_0(\eta)$$

gives $W_S'(\eta) = -\pi f_1(\eta) - (1-\pi)f_0(\eta)$. Since $R_S'(\eta) > 0$, applying the chain rule to $\mathcal{W}_S = W_S \circ R_S^{-1}$ and $\mathcal{U}_S = \mathrm{UWR}_S \circ R_S^{-1}$ gives the two identities. □

We emphasize that the identities in Lemma 2 use only the smoothness and positivity conditions above. The monotone-likelihood-ratio condition is needed for the following convexity argument. Since R_S^{-1} is increasing and $\rho(\eta) = f_0(\eta)/f_1(\eta)$ is nonincreasing, $\rho\left(R_S^{-1}(a)\right)$ is nonincreasing in a. Therefore,

$$\mathcal{W}_S'(a) = -\left[1 + \frac{1-\pi}{\pi}\rho\left(R_S^{-1}(a)\right)\right]$$

is nondecreasing in a, so $\mathcal{W}_S$ is convex on $\mathcal{A}_S$. Similarly,

$$\mathcal{U}_S'(a) = -\frac{1}{\pi}\rho\left(R_S^{-1}(a)\right)$$

is nondecreasing, so $\mathcal{U}_S$ is also convex on $\mathcal{A}_S$.

Lemma 3 (Interior chord bound)**.** Suppose Assumption 1 holds. Let $a_1 < a^\star < a_2$ be points in $\mathcal{A}_S$, and define

$$\bar{\kappa}_{W,S} = \frac{\mathcal{W}_S(a_1) - \mathcal{W}_S(a_2)}{a_2 - a_1}.$$

Then $\mathcal{W}_S(a^\star) - \mathcal{W}_S(a_2) \leq (a_2 - a^\star)\bar{\kappa}_{W,S}$. The same statement holds for $\mathcal{U}_S$, with

$$\bar{\kappa}_{U,S} = \frac{\mathcal{U}_S(a_1) - \mathcal{U}_S(a_2)}{a_2 - a_1}.$$

Proof. Under Assumption 1, $\mathcal{W}_S$ is decreasing and convex, so $|\mathcal{W}_S'(a)|$ is nonincreasing in a. Therefore, the average derivative magnitude over the right subinterval $[a^\star, a_2]$ is no larger than its average over $[a_1, a_2]$:

$$\frac{1}{a_2 - a^\star}\int_{a^\star}^{a_2} |\mathcal{W}_S'(a)|da \leq \frac{1}{a_2 - a_1}\int_{a_1}^{a_2} |\mathcal{W}_S'(a)|da = \bar{\kappa}_{W,S}.$$

Since $\mathcal{W}_S$ is decreasing, we have that $\mathcal{W}_S(a^\star) - \mathcal{W}_S(a_2) = \int_{a^\star}^{a_2} |\mathcal{W}_S'(a)|da$, which gives the result. The argument for $\mathcal{U}_S$ is identical. □

Combining the assumption-free burden decomposition with Lemma 3, if $\delta = a_2 - a^\star$, then

$$\delta \le \mathcal{W}_S(a^\star) - \mathcal{W}_S(a_2) \le \delta \bar{\kappa}_{W,S}.$$

Thus, the lower bound is algebraic and does not depend on Assumption 1, whereas only the chord-based upper bound uses the monotone-likelihood-ratio condition.

Lemma 3 formalizes the following local comparison. For a *fixed* score, a chord measured over a lower-risk portion of its own threshold path upper-bounds the burden reduction obtainable from a small additional relaxation in realized missed-warning risk, provided Assumption 1 continues to hold. Importantly, the exchange rate is score-specific: a chord for one construction cannot in general be used to counterfactually adjust the operating point of another construction.

Numerical application. We then use this calculation only to assess the scale of the warning-burden differences associated with the small realized-risk mismatches discussed in the previous section. On PRM800K ($5 \le K_p \le 20$, minimum reference pool 2), our full model realizes $\mathrm{MWR} = 0.1994$ at target level $\alpha = 0.20$. The two alternative constructions with slightly lower reported warning burden realize $\mathrm{MWR} = 0.2036$ for $\widehat{\lambda}\{f(t) + H\}$ and $\mathrm{MWR} = 0.2029$ for Weibull AFT with H. Their overshoots relative to our realized MWR are therefore $0.2036 - 0.1994 = 0.0042$ and $0.2029 - 0.1994 = 0.0035$, respectively. For each construction, we compute its empirical chord exchange rate along its *own* calibrated threshold path, using the operating points corresponding to the two reported target levels $\alpha = 0.10$ and $\alpha = 0.20$. For $\widehat{\lambda}\{f(t) + H\}$, the corresponding realized risks are 0.1017 and 0.2036, and for Weibull AFT with H they are 0.1017 and 0.2029; hence, in both cases, the matched-risk point 0.1994 lies strictly between the two chord endpoints. The resulting warning-burden chord rates are 3.08 for $\widehat{\lambda}\{f(t) + H\}$ and 3.12 for Weibull AFT with H. Hence, by Lemma 3, tightening each construction from its reported realized risk to the matched-risk point 0.1994 increases its warning burden by at most $0.0042 \times 3.08 = 0.0129$ and $0.0035 \times 3.12 = 0.0109$, respectively. These upper bounds are comparable to the observed warning-rate differences, 0.0109 and 0.0103. Under the empirical admissibility criterion used throughout the experiments, a construction is admissible at target level α only when its realized mean missed-warning rate does not exceed α. Accordingly, $\widehat{\lambda}\{f(t) + H\}$ and Weibull AFT with H are excluded from the matched-constraint comparison at $\alpha = 0.20$ because their realized MWRs, 0.2036 and 0.2029, respectively, exceed the target.

Since these upper bounds are no smaller than the observed differences, the reported burden advantages are of the same order as the burden variation induced by the realized-risk mismatch alone, and therefore do not by themselves evidence lower warning burden at matched missed-warning risk.

B THEORETICAL DETAILS

We first want to note that the guarantee in Proposition 4.1 is marginal over the calibration sample and a new test conversation, conditional on the fitted score-construction pipeline. Since $L_{\mathrm{new}}(\widehat{\eta}) \le \Delta_{\mathrm{new}}$, the nontrivial regime is $\alpha < \pi = \mathbb{P}(\Delta_{\mathrm{new}} = 1)$, in which case, $\mathbb{P}\left(L_{\mathrm{new}}(\widehat{\eta}) = 1 | \Delta_{\mathrm{new}} = 1\right) \le \frac{\alpha}{\pi}$. However, missed-warning validity alone is not enough, since warning at the first turn of every conversation satisfies it at any target level; the criterion cannot separate a useful monitor from one that always warns. We therefore ask whether the propagated feature H allows the calibrated rule to meet the same missed-warning constraint at a lower warning burden. To this end, we generalize the missed-warning loss to an arbitrary pre-response score $S(i,t)$. For such a score and a threshold η, define the turn-level *warning rule*

$$M_{S,i,t}(\eta) = \mathbf{1}\{S(i,t) > \eta\}.$$

The first warning time is $\tau_{S,i}^{\mathrm{warn}}(\eta) = \inf\{t : M_{S,i,t}(\eta) = 1\}$. We define the conversation-level maximum score over the observed at-risk horizon

$$V_{S,i} = \max_{1 \le t \le Y_i} S(i,t), \tag{9}$$

where Y_i is the observed failure or censoring time. Then, the warning-burden functional and the missed-warning loss are

$$W_S(\eta) = \mathbb{P}\{\tau_{S,i}^{\mathrm{warn}}(\eta) \le Y_i\} = \mathbb{P}\{V_{S,i} > \eta\}.$$

$$L_{S,i}(\eta) = \mathbf{1}\{\tau_i \le T_i, \tau_{S,i}^{\mathrm{warn}}(\eta) > \tau_i\} = \Delta_i \mathbf{1}\{V_{S,i} \le \eta\}, \tag{10}$$

and the corresponding missed-warning risk is $R_S(\eta) = \mathbb{E}[L_{S,i}(\eta)] = \mathbb{P}\{\Delta_i = 1, V_{S,i} \le \eta\}$, so that $R_S(\eta) \uparrow \pi$ as $\eta \to \infty$. We then split $\{V_{S,i} > \eta\}$ on Δ_i, which gives $W_S(\eta) = \{\pi - R_S(\eta)\} + \mathbb{P}(\Delta_i = 0, V_{S,i} > \eta)$. Hence, for any feasible threshold satisfying $R_S(\eta) \le \alpha$, $W_S(\eta) \ge \pi - \alpha + \mathbb{P}(\Delta_i = 0, V_{S,i} > \eta) \ge \pi - \alpha$. Thus, minimizing warning burden at level α jointly favors using the available missed-warning budget and reducing unnecessary warnings on non-failed trajectories. If the risk frontier attains $R_S(\eta) = \alpha$, then minimizing $W_S(\eta)$ is equivalent to minimizing the unnecessary-warning probability $\mathbb{P}(\Delta_i = 0, V_{S,i} > \eta)$. For a target missed-warning level $\alpha \in (0, 1)$, the population efficiency objective is

$$\min_{S,\eta} W_S(\eta) \qquad \text{subject to} \qquad R_S(\eta) \le \alpha.$$

Let $X_{i,t} = (t, \widehat{U}_{i,t}, B_{i,t})$ denote the current-turn features excluding propagated history. We compare two score classes, the propagation-oblivious class $\mathcal{S}_0$ and the propagation-aware class $\mathcal{S}_H$, defined by

$$\mathcal{S}_0 = \{S : S(i,t) = s(X_{i,t})\}, \qquad \mathcal{S}_H = \{S : S(i,t) = s(X_{i,t}, H_{i,t})\}. \tag{11}$$

For any score class $\mathcal{S}$, define its oracle warning burden at target level α by

$$W_{\mathcal{S}}^*(\alpha) = \inf_{S \in \mathcal{S},\ \eta : R_S(\eta) \le \alpha} W_S(\eta). \tag{12}$$

In particular, $W_H^*(\alpha) = W_{\mathcal{S}_H}^*(\alpha)$ and $W_0^*(\alpha) = W_{\mathcal{S}_0}^*(\alpha)$. By the nesting above, the infimum over the larger class is no larger, so $W_H^*(\alpha) \le W_0^*(\alpha)$ for every α; strict improvement requires additional structure. This parallels Neyman–Pearson constrained learning, with $R_S(\eta)$ the constrained error and $W_S(\eta)$ the operating cost, though the decision statistic is restricted to the form $V_{S,i} = \max_{t \le Y_i} S(i,t)$.

We will establish an oracle inequality for the fitted rule later. Let $S_H^\dagger \in \mathcal{S}_H$ be the population target of the propagation-aware hazard fit and let $b_H(\alpha) = W_{S_H^\dagger}^*(\alpha) - W_H^*(\alpha) \ge 0$ be its approximation gap relative to the unrestricted propagation-aware oracle; b_H reflects misspecification of the parametric logistic class and does not vanish as n grows.

Assumption 2 (Margin condition)**.** There exists $0 < C_m < \infty$ such that, for all $\eta \in \mathbb{R}$ and all $u \ge 0$,

$$\mathbb{P}\left(|V_{S^*} - \eta| \le u\right) \le C_m u.$$

Assumption 3 (Local threshold regularity)**.** Fix a score S^*, and let η_α^* be a fixed-score oracle threshold satisfying $R_{S^*}(\eta_\alpha^*) = \alpha$ and $W_{S^*}(\eta_\alpha^*) = W_{S^*}^*(\alpha)$. There exists $r_0 > 0$ such that, on the one-sided neighborhood $\mathcal{N}_\alpha = [\eta_\alpha^* - r_0, \eta_\alpha^*]$, the following conditions hold.

1. The missed-warning risk has a local lower slope: there exists $c_R > 0$ such that, for all $\eta \in \mathcal{N}_\alpha$, $R_{S^*}(\eta_\alpha^*) - R_{S^*}(\eta) \ge c_R(\eta_\alpha^* - \eta)$.
2. The grid Γ has one-sided mesh size h_Γ on $\mathcal{N}_\alpha$: for every $\eta \in \mathcal{N}_\alpha$, there exists $\eta_\Gamma \in \Gamma$ such that $\eta - h_\Gamma \le \eta_\Gamma \le \eta$.

We then present our main theorem for the propagation-aware score class $\mathcal{S}_H$, under the NP-style oracle burden $W_H^*(\alpha)$ defined in Eq. equation 12.

Theorem B.1 (Oracle inequality with approximation gap)**.** Let $S_H^\dagger \in \mathcal{S}_H$ denote the population target of the propagation-aware hazard-fitting procedure. For this fixed score, define

$$W_{S_H^\dagger}^*(\alpha) = \inf_{\eta : R_{S_H^\dagger}(\eta) \le \alpha} W_{S_H^\dagger}(\eta),$$

and define the approximation gap $b_H(\alpha) = W_{S_H^\dagger}^*(\alpha) - W_H^*(\alpha)$. Since $S_H^\dagger \in \mathcal{S}_H$, we have $b_H(\alpha) \ge 0$. Suppose Assumption 2 and 3 hold for $S_H^\dagger$ at level $\alpha \in (0, \pi)$, and let $\eta_\alpha^\dagger$ denote the oracle threshold. Let $\widehat{S}_H$ be the fitted propagation-aware score, and let $\Gamma \subset \mathbb{R}$ be a finite threshold grid. Define

$$\widehat{R}_{\mathrm{cal}}(\eta) = \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} \Delta_i \mathbf{1}\{V_{\widehat{S}_H, i} \le \eta\}.$$

Set $\kappa_n = \frac{1-\alpha}{n_{\text{cal}}}$ and $\alpha_n = \alpha - \kappa_n$. Suppose the calibration size satisfies $n_{\text{cal}} \geq \left\lceil \frac{1-\alpha}{\alpha} \right\rceil$. Define the finite-sample corrected calibrated threshold $\widehat{\eta}_H = \max\left\{\eta \in \Gamma : \widehat{R}_{\text{cal}}(\eta) \leq \alpha_n\right\}$. Assume the conditions for Lemmas 9, 10, 11, and 12 hold. Then, with high probability,

$$R_{\widehat{S}_H}(\widehat{\eta}_H) \leq \alpha - \kappa_n + \varepsilon_{\text{cal}},$$
$$W_{\widehat{S}_H}(\widehat{\eta}_H) \leq W_H^*(\alpha) + b_H(\alpha) + O_{\mathbb{P}}\left(\sqrt{\frac{p_H}{n_{\text{tr}}}} + e_K + \sqrt{\frac{\log|\Gamma|}{n_{\text{cal}}}} + \frac{1}{n_{\text{cal}}} + h_\Gamma\right). \tag{13}$$

The generated-feature error vanishes under our construction ($e_K = 0$; Lemma 5).

Proposition B.2 (Oracle dominance)**.** For every $\alpha \in (0,1)$, $W_H^*(\alpha) \leq W_0^*(\alpha)$.

B.1 Proofs for theoretical results

Proof of Proposition 4.1

Proof. Throughout the proof, we condition on the fitted score-construction pipeline, so that the loss functions are exchangeable as assumed. Let $n = n_{\text{cal}}$. Since $R_{i,t} \in (0,1)$ and $\eta_1 < 0$, we have $L_i(\eta_1) = 0$ for every calibration conversation. Hence

$$\frac{n}{n+1}\widehat{R}_{\text{cal}}(\eta_1) + \frac{1}{n+1} = \frac{1}{n+1} \leq \alpha.$$

Thus, the feasible set in Eq. equation 4 is nonempty. Since Γ is finite, $\widehat{\eta}$ is well defined and belongs to Γ. Now augment the calibration sample with the new conversation and define

$$\widehat{R}^+(\eta) = \frac{1}{n+1}\left\{\sum_{i \in \mathcal{I}_{\text{cal}}} L_i(\eta) + L_{\text{new}}(\eta)\right\}.$$

Because $L_{\text{new}}(\eta) \leq 1$, $\widehat{R}^+(\eta) \leq \frac{n}{n+1}\widehat{R}_{\text{cal}}(\eta) + \frac{1}{n+1}$. Therefore, by feasibility of $\widehat{\eta}$, $\widehat{R}^+(\widehat{\eta}) \leq \alpha$. Define the symmetric augmented threshold

$$\widehat{\eta}^+ = \max\left\{\eta \in \Gamma : \widehat{R}^+(\eta) \leq \alpha\right\}.$$

This maximum is well defined because $\widehat{R}^+(\eta_1) = 0$. Moreover, $\widehat{\eta}$ is feasible for the augmented problem, and hence $\widehat{\eta} \leq \widehat{\eta}^+$. By Lemma 4, $L_{\text{new}}(\eta)$ is nondecreasing in η, we have $L_{\text{new}}(\widehat{\eta}) \leq L_{\text{new}}(\widehat{\eta}^+)$. The threshold $\widehat{\eta}^+$ is a permutation-invariant function of the $n+1$ loss functions. By exchangeability,

$$\mathbb{E}\left[L_{\text{new}}(\widehat{\eta}^+)\right] = \mathbb{E}\left[\widehat{R}^+(\widehat{\eta}^+)\right] \leq \alpha,$$

where the inequality follows from the defining feasibility of $\widehat{\eta}^+$. Consequently,

$$\mathbb{E}\left[L_{\text{new}}(\widehat{\eta})\right] \leq \mathbb{E}\left[L_{\text{new}}(\widehat{\eta}^+)\right] \leq \alpha.$$

Finally, since $L_{\text{new}}(\widehat{\eta}) \in \{0,1\}$, $\mathbb{E}\left[L_{\text{new}}(\widehat{\eta})\right] = \mathbb{P}\left\{L_{\text{new}}(\widehat{\eta}) = 1\right\}$, which proves the claim. □

Proof of Theorem B.1

Proof. Let

$$\mathcal{E}_{RW} = \left\{\sup_{\eta\in\Gamma}\left|R_{\widehat{S}_H}(\eta) - R_{S_H^\dagger}(\eta)\right| \leq C_R\varepsilon_S, \quad \sup_{\eta\in\Gamma}\left|W_{\widehat{S}_H}(\eta) - W_{S_H^\dagger}(\eta)\right| \leq C_W\varepsilon_S\right\}$$

be the fitted-to-population risk and burden transfer event from Lemma 8. Let

$$\mathcal{E}_{\text{cal}} = \left\{\sup_{\eta\in\Gamma}\left|\widehat{R}_{\text{cal}}(\eta) - R_{\widehat{S}_H}(\eta)\right| \leq \varepsilon_{\text{cal}}\right\}$$

be the calibration empirical-process event from Lemma 9. By Lemmas 9 and 8, the event $\mathcal{E}_{\text{cal}} \cap \mathcal{E}_{RW}$ holds with high probability. We work on this event. First, by construction, the threshold grid Γ

contains a conservative lower threshold $\eta_{\min}$. Since $\widehat{S}_H$ is a hazard-type score taking values in $[0,1]$, we may choose $\eta_{\min} < 0$, which gives

$$\widehat{R}_{\mathrm{cal}}(\eta_{\min}) = \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} \Delta_i \mathbf{1}\{V_{\widehat{S}_H,i} \le \eta_{\min}\} = 0.$$

Since $n_{\mathrm{cal}} \ge \left\lceil \frac{1-\alpha}{\alpha} \right\rceil$, we have $\alpha_n = \alpha - \frac{1-\alpha}{n_{\mathrm{cal}}} \ge 0$. Hence $\widehat{R}_{\mathrm{cal}}(\eta_{\min}) = 0 \le \alpha_n$, so the corrected calibration feasible set is nonempty and $\widehat{\eta}_H$ is well defined. By Lemma 10, we have

$$R_{\widehat{S}_H}(\widehat{\eta}_H) \le \alpha - \kappa_n + \varepsilon_{\mathrm{cal}}, \quad W_{\widehat{S}_H}(\widehat{\eta}_H) \le W^*_{\widehat{S}_H,\Gamma}(\alpha - \kappa_n - \varepsilon_{\mathrm{cal}}).$$

Set $a = \alpha - \kappa_n - \varepsilon_{\mathrm{cal}}$ and $\xi_n = \kappa_n + \varepsilon_{\mathrm{cal}} + C_R\varepsilon_S$. Then, $a - C_R\varepsilon_S = \alpha - \xi_n$. To apply Lemma 11, we need to verify that the stricter feasible set is nonempty. Under the small-gap and grid-resolution conditions required by Lemma 12, we have $\xi_n \le \alpha$ and $\frac{\xi_n}{c_R} + h_\Gamma \le r_0$. The construction in Lemma 12 first considers the continuous threshold $\widetilde{\eta}_{\xi_n} = \eta^\dagger_\alpha - \frac{\xi_n}{c_R}$, and then, by the one-sided mesh condition, selects a grid point $\eta_{\Gamma,\xi_n} \in \Gamma$ with

$$\widetilde{\eta}_{\xi_n} - h_\Gamma \le \eta_{\Gamma,\xi_n} \le \widetilde{\eta}_{\xi_n}.$$

The same construction gives $R_{S^\dagger_H}(\eta_{\Gamma,\xi_n}) \le \alpha - \xi_n$. Hence

$$\left\{\eta \in \Gamma : R_{S^\dagger_H}(\eta) \le a - C_R\varepsilon_S\right\} = \left\{\eta \in \Gamma : R_{S^\dagger_H}(\eta) \le \alpha - \xi_n\right\} \ne \varnothing.$$

Therefore the feasibility condition required by Lemma 11 is satisfied. Therefore, by Lemma 11, on $\mathcal{E}_{RW}$,

$$W^*_{\widehat{S}_H,\Gamma}(a) \le W^*_{S^\dagger_H,\Gamma}(a - C_R\varepsilon_S) + C_W\varepsilon_S = W^*_{S^\dagger_H,\Gamma}(\alpha - \xi_n) + C_W\varepsilon_S.$$

Combining this with the burden bound from Lemma 10 yields

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W^*_{S^\dagger_H,\Gamma}(\alpha - \xi_n) + C_W\varepsilon_S.$$

Next, by Lemma 12, applied to $S^\dagger_H$ at level α,

$$W^*_{S^\dagger_H,\Gamma}(\alpha - \xi_n) \le W^*_{S^\dagger_H}(\alpha) + C_m\left(\frac{\xi_n}{c_R} + h_\Gamma\right) \implies W_{\widehat{S}_H}(\widehat{\eta}_H) \le W^*_{S^\dagger_H}(\alpha) + C_W\varepsilon_S + C_m\left(\frac{\xi_n}{c_R} + h_\Gamma\right). \tag{14}$$

Using $W^*_{S^\dagger_H}(\alpha) = W^*_H(\alpha) + b_H(\alpha)$, and substituting $\xi_n = \kappa_n + \varepsilon_{\mathrm{cal}} + C_R\varepsilon_S$, we obtain

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W^*_H(\alpha) + b_H(\alpha) + C\left(\varepsilon_S + \varepsilon_{\mathrm{cal}} + \kappa_n + h_\Gamma\right),$$

for a constant $C < \infty$ that absorbs C_W, C_m, c_R^{-1}, and C_R. Finally, by the score perturbation bound and the calibration empirical-process bound, we have

$$\varepsilon_S = O_{\mathbb{P}}\left(\sqrt{\frac{p_H}{n_{\mathrm{tr}}}} + e_K\right), \qquad \varepsilon_{\mathrm{cal}} = O_{\mathbb{P}}\left(\sqrt{\frac{\log|\Gamma|}{n_{\mathrm{cal}}}}\right),$$

and $\kappa_n = \frac{1-\alpha}{n_{\mathrm{cal}}} = O\left(\frac{1}{n_{\mathrm{cal}}}\right)$. Thus,

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W^*_H(\alpha) + b_H(\alpha) + O_{\mathbb{P}}\left(\sqrt{\frac{p_H}{n_{\mathrm{tr}}}} + e_K + \sqrt{\frac{\log|\Gamma|}{n_{\mathrm{cal}}}} + \frac{1}{n_{\mathrm{cal}}} + h_\Gamma\right).$$

This completes the proof. □

Proof of Proposition B.2

Proof. Any $S \in \mathcal{S}_0$ satisfies $S(i,t) = s(X_{i,t}) = s(X_{i,t}, H_{i,t})$ with a map that ignores its last argument, hence $S \in \mathcal{S}_H$ and $\mathcal{S}_0 \subseteq \mathcal{S}_H$. The feasible set in Eq. equation 12 for $\mathcal{S}_H$ therefore contains that for $\mathcal{S}_0$, and the infimum over a larger set is no larger. □

Proof of Corollary 1

Proof. By Theorem B.1, on its high-probability event, there exists a constant $0 < C < \infty$ such that

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W_H^*(\alpha) + b_H(\alpha) + Cr_n^\star.$$

By Definition 1, we have that $W_H^*(\alpha) \le W_0^*(\alpha + \varepsilon_0) - c_{\text{sep}}$. Combining the two displays gives

$$W_{\widehat{S}_H}(\widehat{\eta}_H) \le W_0^*(\alpha + \varepsilon_0) - c_{\text{sep}} + b_H(\alpha) + Cr_n^\star.$$

Hence, if $c_{\text{sep}} > b_H(\alpha) + C(r_n^\star)$, then on the high probability event, the bracket is strictly positive and $W_{\widehat{S}_H}(\widehat{\eta}_H) < W_0^*(\alpha + \varepsilon_0)$. □

B.2 Technical Lemmas and Proofs

Lemma 4 (Monotonicity and right-continuity)**.** For every conversation i, the map $\eta \mapsto L_i(\eta)$ is nondecreasing and right-continuous on $\mathbb{R}$.

Proof. Fix i. For any thresholds $\eta_1 \le \eta_2$,

$$\{t : R_{i,t} > \eta_2\} \subseteq \{t : R_{i,t} > \eta_1\}.$$

Hence $\tau_i^{\text{warn}}(\eta_1) \le \tau_i^{\text{warn}}(\eta_2)$. Therefore the event $\{\tau_i^{\text{warn}}(\eta) > Y_i\}$ is nondecreasing in η. Since

$$L_i(\eta) = \mathbf{1}\{\Delta_i = 1,\ \tau_i^{\text{warn}}(\eta) > Y_i\},$$

it follows that $L_i(\eta)$ is nondecreasing in η. Moreover, on $\{\Delta_i = 1\}$, let $c_i = \max_{1 \le t \le Y_i} R_{i,t}$. Then $L_i(\eta) = \mathbf{1}\{\Delta_i = 1\}\mathbf{1}\{\eta \ge c_i\}$, which is right-continuous in η. On $\{\Delta_i = 0\}$, $L_i(\eta) \equiv 0$. □

To improve readability, we provide the full details for constructing these bounds in Section B.2.1.

Lemma 5 (Preliminary high-probability bounds)**.** Let Assumptions 4–6 hold, and write $\widehat{\theta} = \widehat{\theta}(\widehat{\phi})$ and $\theta^* = \theta^*(\phi^*)$. Then, for every $\delta \in (0, 1)$, there exist $C_\delta < \infty$ and $n_0(\delta)$ such that, for all $n_{\text{tr}} \ge n_0(\delta)$, with $r_{n,\delta} = C_\delta \sqrt{p_H / n_{\text{tr}}}$, the event

$$\mathcal{E}_{\text{tr}} = \left\{\|\widehat{\theta} - \theta^*\|_2 \le r_{n,\delta}\right\} \cap \left\{\widehat{D}_{i,t} = D_{i,t} \text{ for all } (i,t)\right\} \cap \left\{\sup_{i,t} \|D_{i,t}\|_2 \le M_D\right\}$$

has probability at least $1-\delta$. In particular, on $\mathcal{E}_{\text{tr}}$, $\sup_{i,t} \left\|\widehat{D}_{i,t} - D_{i,t}\right\|_2 = 0$, so the generated-feature error in Lemma 6 may be taken as $e_K = 0$. Moreover $\|\widehat{\theta}\|_2 \le M_\theta$ deterministically, since $\widehat{\theta} \in \Theta$ and $\sup_{\theta \in \Theta} \|\theta\|_2 \le M_\theta$.

Proof. By Lemma 13, $\sup_{i,t} \|D_{i,t}\|_2 \le M_D$ holds deterministically. Let $A = \mathcal{E}_{\mathcal{P}} = \{\widehat{\phi} = \phi^*\}$. By Lemma 14 there exists $n_1(\delta)$ with $\mathbb{P}(A) \ge 1 - \delta/2$ for all $n_{\text{tr}} \ge n_1(\delta)$. Define

$$B = \left\{\left\|\widehat{\theta}(\phi^*) - \theta^*(\phi^*)\right\|_2 \le r_{n,\delta}\right\}.$$

Applying Lemma 15 at the fixed, deterministic grid point ϕ^* with confidence level $\delta/2$, and absorbing the corresponding constant into C_δ, there exists $n_2(\delta)$ with $\mathbb{P}(B) \ge 1 - \delta/2$ for all $n_{\text{tr}} \ge n_2(\delta)$. The event B concerns the estimator at the deterministic grid point ϕ^* and is defined irrespective of the selection step; no independence between A and B is therefore required, and the union bound gives

$$\mathbb{P}(A \cap B) \ge 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) \ge 1 - \delta$$

for $n_{\text{tr}} \ge n_0(\delta) := \max\{n_1(\delta), n_2(\delta)\}$. On $A \cap B$ we have $\widehat{\phi} = \phi^*$, hence

$$\widehat{\theta} = \widehat{\theta}(\widehat{\phi}) = \widehat{\theta}(\phi^*), \qquad \widehat{D}_{i,t} = D_{i,t}(\widehat{\phi}) = D_{i,t}(\phi^*) = D_{i,t},$$

so $\|\widehat{\theta} - \theta^*\|_2 \le r_{n,\delta}$ and $\sup_{i,t} \|\widehat{D}_{i,t} - D_{i,t}\|_2 = 0$. Together with the deterministic design bound this proves the claim. Finally $\widehat{\theta} \in \Theta$ by construction, and Assumption 5 gives $\sup_{\theta \in \Theta} \|\theta\|_2 \le M_\theta$, so $\|\widehat{\theta}\|_2 \le M_\theta$. □

Lemma 6 (Uniform score perturbation)**.** Let $D_{i,t}$ denote the oracle pre-response design vector for the propagation-aware hazard score, as defined above, and let $\widehat{D}_{i,t}$ be its estimated counterpart. We define

$$S_\theta(i,t) = g(\theta^\top D_{i,t}), \qquad \widehat{S}(i,t) = g(\widehat{\theta}^\top \widehat{D}_{i,t}),$$

where $g : \mathbb{R} \to \mathbb{R}$ is L_g-Lipschitz. Then, on the event $\mathcal{E}_{\mathrm{tr}}$ of Lemma 5, which holds with probability at least $1 - \delta$,

$$\sup_{i,t} \left| \widehat{S}(i,t) - S_{\theta^*}(i,t) \right| \le L_g \left(M_D r_{n,\delta} + M_\theta e_K \right).$$

Since $e_K = 0$ on that event and $r_{n,\delta} = C_\delta \sqrt{p_H / n_{\mathrm{tr}}}$, this gives

$$\sup_{i,t} \left| \widehat{S}(i,t) - S_{\theta^*}(i,t) \right| = O_{\mathbb{P}} \left(\sqrt{\frac{p_H}{n_{\mathrm{tr}}}} \right).$$

Proof. For any at-risk pair (i,t), by Lipschitz continuity of g,

$$\left| \widehat{S}(i,t) - S_{\theta^*}(i,t) \right| = \left| g(\widehat{\theta}^\top \widehat{D}_{i,t}) - g((\theta^*)^\top D_{i,t}) \right| \le L_g \left| \widehat{\theta}^\top \widehat{D}_{i,t} - (\theta^*)^\top D_{i,t} \right|.$$

Decompose the linear predictor difference as

$$\begin{aligned} &\widehat{\theta}^\top \widehat{D}_{i,t} - (\theta^*)^\top D_{i,t} = (\widehat{\theta} - \theta^*)^\top D_{i,t} + \widehat{\theta}^\top (\widehat{D}_{i,t} - D_{i,t}). \\ &\implies \left| \widehat{\theta}^\top \widehat{D}_{i,t} - (\theta^*)^\top D_{i,t} \right| \le \|\widehat{\theta} - \theta^*\|_2 \|D_{i,t}\|_2 + \|\widehat{\theta}\|_2 \|\widehat{D}_{i,t} - D_{i,t}\|_2. \end{aligned} \tag{15}$$

Taking the supremum over all at-risk pairs (i,t) gives

$$\sup_{i,t} \left| \widehat{S}(i,t) - S_{\theta^*}(i,t) \right| \le L_g \left[\|\widehat{\theta} - \theta^*\|_2 \sup_{i,t} \|D_{i,t}\|_2 + \|\widehat{\theta}\|_2 \sup_{i,t} \|\widehat{D}_{i,t} - D_{i,t}\|_2 \right].$$

On the event $\mathcal{E}_{\mathrm{tr}}$ from Lemma 5, we obtain

$$\sup_{i,t} \left| \widehat{S}(i,t) - S_{\theta^*}(i,t) \right| \le L_g (M_D r_{n,\delta} + M_\theta e_K),$$

and $e_K = 0$ on that event, so the final stochastic order follows from $r_{n,\delta} = C_\delta \sqrt{p_H / n_{\mathrm{tr}}}$. □

Lemma 7 (Max-score perturbation)**.** For a score function S, let the conversation-level at-risk maximum score by Eq.equation 9, and let $V_{\widehat{S},i} = \max_{1 \le t \le Y_i} \widehat{S}(i,t)$ and $V_{S^*,i} = \max_{1 \le t \le Y_i} S^*(i,t)$. Then, the following deterministic inequality holds:

$$\sup_i \left| V_{\widehat{S},i} - V_{S^*,i} \right| \le \sup_{i,t:1 \le t \le Y_i} \left| \widehat{S}(i,t) - S^*(i,t) \right|.$$

Consequently, under the conditions of Lemma 6, let $\varepsilon_S = L_g M_D r_{n,\delta}$. Then, with the probability at least $1 - \delta$, we have

$$\sup_i \left| V_{\widehat{S},i} - V_{S^*,i} \right| \le \varepsilon_S.$$

Proof. Fix a conversation i. Let

$$\widehat{t}_i \in \arg\max_{1 \le t \le Y_i} \widehat{S}(i,t), \qquad t_i^* \in \arg\max_{1 \le t \le Y_i} S^*(i,t).$$

Since t_i^* maximizes $S^*(i,t)$, we have $S^*(i,\widehat{t}_i) \le V_{S^*,i}$. Therefore,

$$V_{\widehat{S},i} - V_{S^*,i} = \widehat{S}(i,\widehat{t}_i) - V_{S^*,i} \le \widehat{S}(i,\widehat{t}_i) - S^*(i,\widehat{t}_i) \le \sup_{1 \le t \le Y_i} \left| \widehat{S}(i,t) - S^*(i,t) \right|. \tag{16}$$

Similarly, since $\widehat{t}_i$ maximizes $\widehat{S}(i,t)$, we have $\widehat{S}(i,t_i^*) \le V_{\widehat{S},i}$. Hence

$$V_{S^*,i} - V_{\widehat{S},i} = S^*(i,t_i^*) - V_{\widehat{S},i} \le S^*(i,t_i^*) - \widehat{S}(i,t_i^*) \le \sup_{1 \le t \le Y_i} \left| \widehat{S}(i,t) - S^*(i,t) \right|. \tag{17}$$

Combining the two inequalities and taking the supremum over i yields

$$\sup_i \left| V_{\widehat{S},i} - V_{S^*,i} \right| \le \sup_{i,t:1 \le t \le Y_i} \left| \widehat{S}(i,t) - S^*(i,t) \right|.$$

The high-probability statement follows immediately by applying this deterministic inequality on the event from Lemma 6. □

The preceding perturbation bounds are deterministic conditional on the high-probability estimation event. Therefore, whenever the uniform score perturbation bound holds over a collection of at-risk trajectories, the corresponding max-score perturbation bound holds over the same collection. In particular, for population warning burden and missed-warning risk, we apply the same argument to an independent generic test conversation drawn from the same population, provided that the score-perturbation event is stated uniformly over the population support of at-risk pre-response features (so that, conditional on the high-probability estimation event, the perturbation event holds surely for any test conversation drawn from the population).

Lemma 8 (Margin transfer for warning burden and missed-warning risk)**.** Let $V_{\widehat{S}}$ and V_{S^*} denote the conversation-level at-risk maximum scores for a generic conversation under the fitted score $\widehat{S}$ and the oracle score S^*, respectively. Define $W_{\widehat{S}}(\eta) = \mathbb{P}(V_{\widehat{S}} > \eta)$, $W_{S^*}(\eta) = \mathbb{P}(V_{S^*} > \eta)$, $R_{\widehat{S}}(\eta) = \mathbb{P}(\Delta = 1, V_{\widehat{S}} \le \eta)$, and $R_{S^*}(\eta) = \mathbb{P}(\Delta = 1, V_{S^*} \le \eta)$. Let $\mathcal{E}_{\mathrm{tr}}$ denote the high-probability estimation event from Lemma 5. On $\mathcal{E}_{\mathrm{tr}}$, the score perturbation bound holds uniformly over the relevant population support of at-risk pre-response features. Hence, for the generic test conversation, $\mathcal{E}_S = \left\{|V_{\widehat{S}} - V_{S^*}| \le \varepsilon_S\right\}$ holds almost surely conditional on the training data and $\mathcal{E}_{\mathrm{tr}}$. If Assumption 2 holds, then, on $\mathcal{E}_{\mathrm{tr}}$, we have

$$\sup_{\eta\in\Gamma}\left|W_{\widehat{S}}(\eta) - W_{S^*}(\eta)\right| \le C_m\varepsilon_S, \qquad \sup_{\eta\in\Gamma}\left|R_{\widehat{S}}(\eta) - R_{S^*}(\eta)\right| \le C_m\varepsilon_S.$$

Proof. Work on $\mathcal{E}_{\mathrm{tr}}$. Conditional on $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{E}_{\mathrm{tr}}$, the fitted score $\widehat{S}$ is fixed, ε_S is a fixed constant, and the event $\mathcal{E}_S = \{|V_{\widehat{S}} - V_{S^*}| \le \varepsilon_S\}$ holds almost surely for the generic test conversation. Fix $\eta \in \Gamma$. On the event $\mathcal{E}_S$, if $V_{\widehat{S}} > \eta$ and $V_{S^*} \le \eta$, then $V_{S^*} > \eta - \varepsilon_S$. Therefore,

$$\{V_{\widehat{S}} > \eta,\ V_{S^*} \le \eta\} \subseteq \{\eta - \varepsilon_S < V_{S^*} \le \eta\} \subseteq \{|V_{S^*} - \eta| \le \varepsilon_S\}.$$

Similarly, if $V_{S^*} > \eta$ and $V_{\widehat{S}} \le \eta$, then $V_{S^*} \le \eta + \varepsilon_S$, so

$$\{V_{S^*} > \eta,\ V_{\widehat{S}} \le \eta\} \subseteq \{\eta < V_{S^*} \le \eta + \varepsilon_S\} \subseteq \{|V_{S^*} - \eta| \le \varepsilon_S\}.$$

Hence the two events $\{V_{\widehat{S}} > \eta\}$ and $\{V_{S^*} > \eta\}$ can differ only when V_{S^*} lies within ε_S of the threshold η. Since the generic test conversation is independent of the training data, conditioning on $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{E}_{\mathrm{tr}}$ does not change the population law of V_{S^*}. Therefore, Assumption 2 applies with the deterministic radius $u = \varepsilon_S$, giving

$$\left|W_{\widehat{S}}(\eta) - W_{S^*}(\eta)\right| \le \mathbb{P}\left(|V_{S^*} - \eta| \le \varepsilon_S\right) \le C_m\varepsilon_S.$$

For the missed-warning risk, note that

$$\left|\mathbf{1}\{\Delta = 1, V_{\widehat{S}} \le \eta\} - \mathbf{1}\{\Delta = 1, V_{S^*} \le \eta\}\right| \le \mathbf{1}\{|V_{S^*} - \eta| \le \varepsilon_S\}.$$

Taking expectations and applying the same margin step as above gives

$$\left|R_{\widehat{S}}(\eta) - R_{S^*}(\eta)\right| \le \mathbb{P}\left(|V_{S^*} - \eta| \le \varepsilon_S\right) \le C_m\varepsilon_S.$$

Since the argument holds for each $\eta \in \Gamma$, taking the supremum over $\eta \in \Gamma$ proves the uniform bounds. □

We emphasize that the observational unit in our analysis is a complete conversation. Thus, the sample is organized at the conversation level: under a fixed conversation-generation protocol, distinct conversations are regarded as separate observational units drawn from the target distribution over conversation instances. With the random 60/20/20 split performed at the conversation level, and conditional on the training set and the fitted score $\widehat{S}$, the calibration and test conversations can be treated as iid samples from the same target conversation population. Let $\mathcal{I}_{\mathrm{cal}}$ denote the calibration set with $|\mathcal{I}_{\mathrm{cal}}| = n_{\mathrm{cal}}$. Suppose the fitted score $\widehat{S}$ is trained using data independent of the calibration set and is fixed before calibration. Then, we can have the following lemma

Lemma 9 (Calibration empirical-process bound)**.** Let Γ be a finite threshold grid. For each $\eta \in \Gamma$, let the calibration missed-warning loss, $L_i(\eta)$, for each $i \in \mathcal{I}_{\mathrm{cal}}$ be given by Eq. equation 10 and the empirical calibration risk

$$\widehat{R}_{\mathrm{cal}}(\eta) = \frac{1}{n_{\mathrm{cal}}}\sum_{i\in\mathcal{I}_{\mathrm{cal}}} L_i(\eta).$$

Let $R_{\widehat{S}}(\eta) = \mathbb{E}\{L_i(\eta)\} = \mathbb{P}(\Delta = 1, V_{\widehat{S}} \le \eta)$ denote the corresponding population missed-warning risk, conditional on the trained score $\widehat{S}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{\eta\in\Gamma} \left|\widehat{R}_{\mathrm{cal}}(\eta) - R_{\widehat{S}}(\eta)\right| \le \sqrt{\frac{\log(2|\Gamma|/\delta)}{2n_{\mathrm{cal}}}}.$$

Equivalently, $\sup_{\eta\in\Gamma} \left|\widehat{R}_{\mathrm{cal}}(\eta) - R_{\widehat{S}}(\eta)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log|\Gamma|}{n_{\mathrm{cal}}}}\right)$.

Proof. Condition on the training data used to construct $\widehat{S}$. Then $\widehat{S}$ is fixed with respect to the calibration set. For each fixed $\eta \in \Gamma$, the variables $L_i(\eta) = \Delta_i \mathbf{1}\{V_{\widehat{S},i} \le \eta\}$, for each $i \in \mathcal{I}_{\mathrm{cal}}$, are independent, identically distributed, and bounded in $[0,1]$. Moreover,

$$\mathbb{E}\{L_i(\eta) \mid \mathcal{D}_{\mathrm{tr}}\} = R_{\widehat{S}}(\eta).$$

By Hoeffding's inequality, for any $u > 0$,

$$\mathbb{P}\left(\left|\widehat{R}_{\mathrm{cal}}(\eta) - R_{\widehat{S}}(\eta)\right| > u \middle| \mathcal{D}_{\mathrm{tr}}\right) \le 2\exp(-2n_{\mathrm{cal}}u^2).$$

Taking a union bound over the finite threshold grid Γ, we obtain

$$\mathbb{P}\left(\sup_{\eta\in\Gamma}\left|\widehat{R}_{\mathrm{cal}}(\eta) - R_{\widehat{S}}(\eta)\right| > u \middle| \mathcal{D}_{\mathrm{tr}}\right) \le 2|\Gamma|\exp(-2n_{\mathrm{cal}}u^2).$$

Choosing $u = \sqrt{\frac{\log(2|\Gamma|/\delta)}{2n_{\mathrm{cal}}}}$ makes the right-hand side equal to δ. Therefore, conditional on the training data, the desired bound holds with probability at least $1-\delta$. Removing the conditioning gives the same unconditional probability bound. □

Lemma 10 (Calibration threshold comparison with finite-sample correction)**.** Let $\Gamma \subset \mathbb{R}$ be a finite threshold grid. For a fitted score $\widehat{S}$, define $W_{\widehat{S}}(\eta) = \mathbb{P}(V_{\widehat{S}} > \eta)$. Let $\widehat{R}_{\mathrm{cal}}(\eta)$ and $R_{\widehat{S}}(\eta)$ from Lemma 9. Define $\kappa_n = \frac{1-\alpha}{n_{\mathrm{cal}}}$ and $\alpha_n = \alpha - \kappa_n$. The finite-sample corrected calibration rule is

$$\frac{n_{\mathrm{cal}}}{n_{\mathrm{cal}}+1}\widehat{R}_{\mathrm{cal}}(\eta) + \frac{1}{n_{\mathrm{cal}}+1} \le \alpha.$$

equivalently, $\widehat{\eta} = \max\left\{\eta \in \Gamma : \widehat{R}_{\mathrm{cal}}(\eta) \le \alpha_n\right\}$. Assume the feasible set is nonempty; in particular, this requires $\alpha_n \ge 0$, equivalently $n_{\mathrm{cal}} \ge \frac{1-\alpha}{\alpha}$. Let

$$\mathcal{E}_{\mathrm{cal}} = \left\{\sup_{\eta\in\Gamma}\left|\widehat{R}_{\mathrm{cal}}(\eta) - R_{\widehat{S}}(\eta)\right| \le \varepsilon_{\mathrm{cal}}\right\}.$$

On the event $\mathcal{E}_{\mathrm{cal}}$, we have that $R_{\widehat{S}}(\widehat{\eta}) \le \alpha - \kappa_n + \varepsilon_{\mathrm{cal}}$. Moreover, define the grid oracle burden

$$W^*_{\widehat{S},\Gamma}(a) = \inf_{\eta\in\Gamma: R_{\widehat{S}}(\eta)\le a} W_{\widehat{S}}(\eta), \tag{18}$$

with the convention that the infimum over an empty feasible set is $+\infty$. Then

$$W_{\widehat{S}}(\widehat{\eta}) \le W^*_{\widehat{S},\Gamma}\left(\alpha - \kappa_n - \varepsilon_{\mathrm{cal}}\right).$$

Proof. On $\mathcal{E}_{\mathrm{cal}}$, the selected threshold $\widehat{\eta}$ satisfies $\widehat{R}_{\mathrm{cal}}(\widehat{\eta}) \le \alpha_n = \alpha - \kappa_n$. Therefore,

$$R_{\widehat{S}}(\widehat{\eta}) \le \widehat{R}_{\mathrm{cal}}(\widehat{\eta}) + \varepsilon_{\mathrm{cal}} \le \alpha - \kappa_n + \varepsilon_{\mathrm{cal}}.$$

This proves the approximate feasibility claim. Next, take any $\eta \in \Gamma$ such that $R_{\widehat{S}}(\eta) \le \alpha - \kappa_n - \varepsilon_{\mathrm{cal}}$. Then, on $\mathcal{E}_{\mathrm{cal}}$,

$$\widehat{R}_{\mathrm{cal}}(\eta) \le R_{\widehat{S}}(\eta) + \varepsilon_{\mathrm{cal}} \le \alpha - \kappa_n = \alpha_n.$$

Thus η is empirically feasible for the corrected calibration rule. Since $\widehat{\eta}$ is the largest empirically feasible threshold, $\widehat{\eta} \ge \eta$. Since $W_{\widehat{S}}(\eta) = \mathbb{P}(V_{\widehat{S}} > \eta)$ is nonincreasing in η, we have $W_{\widehat{S}}(\widehat{\eta}) \le W_{\widehat{S}}(\eta)$. Taking the infimum over all $\eta \in \Gamma$ satisfying $R_{\widehat{S}}(\eta) \le \alpha - \kappa_n - \varepsilon_{\mathrm{cal}}$ gives

$$W_{\widehat{S}}(\widehat{\eta}) \le W^*_{\widehat{S},\Gamma}\left(\alpha - \kappa_n - \varepsilon_{\mathrm{cal}}\right).$$

This proves the result. □

Lemma 11 (Fitted-grid oracle transfer)**.** For any score S, let the grid-restricted oracle burden at risk level a be defined by Eq. equation 18. Let

$$\mathcal{E}_{RW} = \left\{ \sup_{\eta\in\Gamma} \left| R_{\widehat{S}}(\eta) - R_{S^*}(\eta) \right| \le C_R \varepsilon_S \right\} \cap \left\{ \sup_{\eta\in\Gamma} \left| W_{\widehat{S}}(\eta) - W_{S^*}(\eta) \right| \le C_W \varepsilon_S \right\}.$$

By Lemma 8, $\mathcal{E}_{RW}$ holds with the same high probability as the score-perturbation event. Then, on $\mathcal{E}_{RW}$, for any a such that the stricter feasible set $\{\eta \in \Gamma : R_{S^*}(\eta) \le a - C_R\varepsilon_S\}$ is nonempty,

$$W^*_{\widehat{S},\Gamma}(a) \le W^*_{S^*,\Gamma}(a - C_R\varepsilon_S) + C_W\varepsilon_S.$$

Proof. Fix a. If the feasible set $\{\eta \in \Gamma : R_{S^*}(\eta) \le a - C_R\varepsilon_S\}$ is empty, then $W^*_{S^*,\Gamma}(a - C_R\varepsilon_S) = +\infty$ by convention, and the desired inequality is trivial. Otherwise, the feasible set is nonempty. Since Γ is finite, there exists $\eta^* \in \Gamma$ attaining

$$W^*_{S^*,\Gamma}(a - C_R\varepsilon_S) = W_{S^*}(\eta^*)$$

subject to $R_{S^*}(\eta^*) \le a - C_R\varepsilon_S$. On $\mathcal{E}_{RW}$, the risk perturbation bound gives

$$R_{\widehat{S}}(\eta^*) \le R_{S^*}(\eta^*) + C_R\varepsilon_S \le a.$$

Therefore, η^* is feasible for the fitted grid oracle $W^*_{\widehat{S},\Gamma}(a)$. Hence $W^*_{\widehat{S},\Gamma}(a) \le W_{\widehat{S}}(\eta^*)$. Using the burden perturbation bound on $\mathcal{E}_{RW}$, we have $W_{\widehat{S}}(\eta^*) \le W_{S^*}(\eta^*) + C_W\varepsilon_S$. Combining the previous two displays yields

$$W^*_{\widehat{S},\Gamma}(a) \le W_{S^*}(\eta^*) + C_W\varepsilon_S = W^*_{S^*,\Gamma}(a - C_R\varepsilon_S) + C_W\varepsilon_S.$$

This proves the claim. □

Lemma 12 (Grid approximation)**.** Fix a score S^*. For any risk level a, define the continuous-threshold oracle burden and the grid-restricted oracle burden

$$W^*_{S^*}(a) = \inf_{\eta: R_{S^*}(\eta)\le a} W_{S^*}(\eta), \qquad W^*_{S^*,\Gamma}(a) = \inf_{\eta\in\Gamma: R_{S^*}(\eta)\le a} W_{S^*}(\eta).$$

Let η^*_α be an oracle threshold, and suppose Assumption 2 and 3 hold at η^*_α. Then, for every $0 \le \xi \le \alpha$ satisfying $\frac{\xi}{c_R} + h_\Gamma \le r_0$, we have

$$W^*_{S^*,\Gamma}(\alpha - \xi) \le W^*_{S^*}(\alpha) + C_m \left(\frac{\xi}{c_R} + h_\Gamma \right).$$

In particular, $W^*_{S^*,\Gamma}(\alpha-\xi) \le W^*_{S^*}(\alpha) + C_\Gamma(\xi + h_\Gamma)$, where one may take $C_\Gamma = C_m \max\{c_R^{-1}, 1\}$.

Proof. Fix $\xi \ge 0$ such that $\xi/c_R + h_\Gamma \le r_0$. Define the conservative continuous threshold $\widetilde{\eta}_\xi = \eta^*_\alpha - \frac{\xi}{c_R}$. Because $\xi/c_R \le r_0$, we have $\widetilde{\eta}_\xi \in \mathcal{N}_\alpha$. By the local lower-slope condition for R_{S^*},

$$R_{S^*}(\eta^*_\alpha) - R_{S^*}(\widetilde{\eta}_\xi) \ge c_R(\eta^*_\alpha - \widetilde{\eta}_\xi) = \xi.$$

Since $R_{S^*}(\eta^*_\alpha) = \alpha$, this implies $R_{S^*}(\widetilde{\eta}_\xi) \le \alpha - \xi$. By the one-sided grid condition, there exists $\eta_\Gamma \in \Gamma$ such that $\widetilde{\eta}_\xi - h_\Gamma \le \eta_\Gamma \le \widetilde{\eta}_\xi$. Moreover,

$$\eta^*_\alpha - \eta_\Gamma \le \eta^*_\alpha - \widetilde{\eta}_\xi + h_\Gamma = \frac{\xi}{c_R} + h_\Gamma \le r_0, \tag{19}$$

so $\eta_\Gamma \in \mathcal{N}_\alpha$. Because $R_{S^*}(\eta)$ is nondecreasing in η, and because $\eta_\Gamma \le \widetilde{\eta}_\xi$, $R_{S^*}(\eta_\Gamma) \le R_{S^*}(\widetilde{\eta}_\xi) \le \alpha - \xi$. Therefore, η_Γ is feasible for the grid oracle $W^*_{S^*,\Gamma}(\alpha-\xi)$. Hence $W^*_{S^*,\Gamma}(\alpha-\xi) \le W_{S^*}(\eta_\Gamma)$. By Assumption 2, since $\eta_\Gamma \le \eta^*_\alpha$,

$$W_{S^*}(\eta_\Gamma) - W_{S^*}(\eta^*_\alpha) = \mathbb{P}(\eta_\Gamma < V_{S^*} \le \eta^*_\alpha) \le \mathbb{P}\left(|V_{S^*} - \eta^*_\alpha| \le \eta^*_\alpha - \eta_\Gamma\right) \le C_m(\eta^*_\alpha - \eta_\Gamma).$$

Furthermore, by Eq. equation 19 and combining the previous displays gives

$$W^*_{S^*,\Gamma}(\alpha - \xi) \le W_{S^*}(\eta^*_\alpha) + C_m \left(\frac{\xi}{c_R} + h_\Gamma \right).$$

Finally, since $W_{S^*}(\eta^*_\alpha) = W^*_{S^*}(\alpha)$, we obtain $W^*_{S^*,\Gamma}(\alpha - \xi) \le W^*_{S^*}(\alpha) + C_m \left(\frac{\xi}{c_R} + h_\Gamma \right)$. The final bound follows by taking $C_\Gamma = C_m \max\{c_R^{-1}, 1\}$. □

B.2.1 LEMMAS ON HIGH-PROBABILITY BOUNDS

This subsection provides the technical details for the ingredients used in Lemma 5, including a uniform bound on the pre-response design vectors and convergence rates for the kernel-selection and hazard-estimation steps. Throughout, K_f, $\dim(B)$, and hence p_H, are fixed, so all rates below are finite-dimensional. We write $\mathcal{I}_{\mathrm{fit}} \subseteq \mathcal{I}_{\mathrm{train}}$ for the subset of training units used to fit the hazard model, with $n_{\mathrm{tr}} = |\mathcal{I}_{\mathrm{fit}}|$. The fitting and validation subsets are taken to be fixed positive fractions of $\mathcal{I}_{\mathrm{train}}$, so both have size of order $|\mathcal{I}_{\mathrm{train}}|$. The independent sampling unit is a conversation in the multi-turn setting and a problem p in the chain-of-thought setting, where all chains from the same problem are treated as a single cluster. Accordingly, n_{tr} counts independent sampling units rather than at-risk turns. To avoid collision with the first-failure event label $Z_{i,t} = \mathbf{1}\{Y_i = t, \Delta_i = 1\}$, we write the pre-response design vector at (i,t) under kernel parameter ϕ as

$$D_{i,t}(\phi) = \left(1,\ \psi_1(t), \ldots, \psi_{K_f}(t), \widehat{U}_{i,t}, H_{i,t}(\phi), B_{i,t}^\top\right)^\top \in \mathbb{R}^{p_H}, \qquad p_H = 3 + K_f + \dim(B),$$

so that $\lambda_\theta(i,t) = \sigma\{\theta^\top D_{i,t}(\phi)\}$ with $\theta = (\beta_0, \gamma^\top, \beta_1, \alpha, \beta_3^\top)^\top$; the three scalar coordinates are the intercept and the coefficients on $\widehat{U}_{i,t}$ and $H_{i,t}(\phi)$. We abbreviate $D_{i,t} = D_{i,t}(\phi^*)$ and $\widehat{D}_{i,t} = D_{i,t}(\widehat{\phi})$, and write $\mathcal{P} \subset [\underline{\phi}, \bar{\phi}] \subset (0,1)$ for the fixed kernel grid, reserving Φ for the standard normal distribution function.

Assumption 4 (Sampling, bounded design, and fixed feature pipeline)**.** The diagnostic sampling rule and semantic feature-construction procedure are fixed before hazard fitting. In particular, the pilot scoring parameters $\widehat{\beta}^{(0)}$, policy threshold η_{pol}, allocation map g, and semantic clustering rule are either specified a priori or estimated using data disjoint from $\mathcal{I}_{\mathrm{fit}}$. Conditional on this fixed upstream pipeline, the sampling units in $\mathcal{I}_{\mathrm{fit}}$ are i.i.d. draws from the target population. Each sampling unit contributes a uniformly bounded number of at-risk observations. In the multi-turn setting, $T_i \le T_{\max} < \infty$ and the diagnostic budget satisfies $K_{i,t} \le \bar{K} = K_{\min} + B_{\max}$. In the chain-of-thought setting, $T_{p,k} \le T_{\max}$ and $K_p \le K_{\max} < \infty$, so each problem-level unit contributes at most $N_{\max} = K_{\max} T_{\max}$ at-risk terms. The components of the design vector are also uniformly bounded: the turn basis satisfies $\max_k \sup_{t \le T_{\max}} |\psi_k(t)| \le M_\psi$, and the auxiliary covariates satisfy $\sup_{i,t} \|B_{i,t}\|_2 \le M_B$.

In the multi-turn experiments, the diagnostic budget is fixed at $K_{i,t} \equiv K$, and the semantic feature-construction procedure is fixed before hazard fitting, so Assumption 4 holds directly. In the chain-of-thought setting, no learned allocation rule is used. Because the leave-one-out construction induces dependence among sibling chains from the same problem, we treat the problem p, rather than an individual chain (p,k), as the independent sampling unit, consistent with the problem-level splits used in the experiments. The bounded-cluster condition $K_p \le K_{\max}$ is satisfied in the capped regime $5 \le K_p \le 20$ considered in the theory; the unrestricted regime is excluded and is discussed separately in Appendix C.

Assumption 5 (Hazard regularity)**.** For each $\phi \in \mathcal{P}$, the hazard parameter is estimated by

$$\widehat{\theta}(\phi) = \arg\min_{\theta \in \Theta} \widehat{L}_n(\theta; \phi), \qquad \widehat{L}_n(\theta; \phi) = \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathcal{I}_{\mathrm{fit}}} m_{\theta,\phi}(\mathcal{T}_i) + \lambda_n \omega(\theta),$$

where $\lambda_n = O(n_{\mathrm{tr}}^{-1/2})$. The penalty ω is convex and L_ω-Lipschitz on a compact convex parameter space $\Theta \subset \mathbb{R}^{p_H}$ satisfying $\sup_{\theta \in \Theta} \|\theta\|_2 \le M_\theta$. For each $\phi \in \mathcal{P}$, the population criterion $L(\theta; \phi) = \mathbb{E}\left[m_{\theta,\phi}(\mathcal{T}_i)\right]$ has a unique minimizer $\theta^*(\phi)$ in the interior of Θ. Moreover, the population second-moment matrix $\Sigma(\phi) = \mathbb{E}\left[\sum_{j \in \mathcal{J}_i} D_{i,j}(\phi) D_{i,j}(\phi)^\top\right]$ satisfies $\lambda_{\min}\{\Sigma(\phi)\} \ge c_D > 0$ uniformly over $\phi \in \mathcal{P}$. Here $\mathcal{J}_i$ denotes the set of at-risk observations contributed by sampling unit i. In the multi-turn setting, $\mathcal{J}_i = \{t : 1 \le t \le Y_i\}$; in the chain-of-thought setting, with $i = p$, $\mathcal{J}_i$ contains all eligible chain-step pairs (k,t) from problem p.

Assumption 6 (Kernel separation on the grid)**.** The kernel parameter is selected by $\widehat{\phi} \in \arg\max_{\phi \in \mathcal{P}} \widehat{Q}_n(\phi)$, where $\widehat{Q}_n$ is evaluated on the held-out validation subset of $\mathcal{I}_{\mathrm{train}}$. Let Q denote its population counterpart, and assume $\sup_{\phi \in \mathcal{P}} \left|\widehat{Q}_n(\phi) - Q(\phi)\right| = O_{\mathbb{P}}\left(n_{\mathrm{tr}}^{-1/2}\right)$. The population criterion has a unique maximizer ϕ^* on the fixed grid $\mathcal{P}$, with strictly positive separation $\Delta_{\mathcal{P}} = \min_{\phi \in \mathcal{P} \setminus \{\phi^*\}} \{Q(\phi^*) - Q(\phi)\} > 0$.

For a fixed finite grid, the uniform rate above follows from standard concentration of the validation criterion together with the root-n_{tr} estimation rate of the fitted candidate models. We state the rate directly because $\widehat{Q}_n(\phi)$ depends on ϕ through fitted model parameters, so its deviation from $Q(\phi)$ reflects both validation-sample variation and parameter-estimation error.

Lemma 13 (Uniform design bound)**.** Under Assumption 4, for every $\phi \in [\underline{\phi}, \bar{\phi}]$, $\sup_{i,j} \|D_{i,j}(\phi)\|_2 \le M_D$, where $M_D := \left\{2 + K_f M_\psi^2 + \left(\log \bar{S} + \frac{1}{2}\right)^2 + M_B^2\right\}^{1/2}$, where $\bar{S} := \max\{\bar{K}, K_{\max} - 1\}$.

Proof. The intercept contributes 1, while the turn basis contributes at most $K_f M_\psi^2$. Let $S_{i,j}$ denote the effective semantic sample size. $S_{i,j} = K_{i,t}$ in the multi-turn setting and $S_{i,j} = n_{p,k,t}$ in the chain-of-thought setting. By Assumption 4, $S_{i,j} \le \bar{S}$. The Miller–Madow corrected dispersion score therefore satisfies $0 \le \widehat{U}_{i,j} \le \log m_{i,j} + \frac{m_{i,j}-1}{2S_{i,j}} \le \log \bar{S} + \frac{1}{2}$. The propagation coordinate satisfies $H_{i,j}(\phi) \in [0,1]$, since the historical weights are nonnegative and sum to one and $\widehat{q}_{i,j} \in [0,1]$. Finally, $\|B_{i,j}\|_2 \le M_B$. Summing the squared contributions gives the stated bound. □

Lemma 14 (Kernel selection consistency)**.** Under Assumption 6, $\mathbb{P}(\widehat{\phi} \neq \phi^*) \to 0$. Moreover, on the event $\mathcal{E}_{\mathcal{P}} = \{\widehat{\phi} = \phi^*\}$, one has $\widehat{D}_{i,j} = D_{i,j}$ for every at-risk observation (i,j).

Proof. If $\widehat{\phi} \neq \phi^*$, then $\widehat{Q}_n(\widehat{\phi}) \ge \widehat{Q}_n(\phi^*)$ while $Q(\phi^*) - Q(\widehat{\phi}) \ge \Delta_{\mathcal{P}}$. Hence, $\sup_{\phi \in \mathcal{P}} \left|\widehat{Q}_n(\phi) - Q(\phi)\right| \ge \frac{\Delta_{\mathcal{P}}}{2}$. The probability of this event tends to zero by Assumption 6. On $\mathcal{E}_{\mathcal{P}}$, $\widehat{\phi} = \phi^*$, and therefore $\widehat{D}_{i,j} = D_{i,j}(\widehat{\phi}) = D_{i,j}(\phi^*) = D_{i,j}$. □

Remark 2 (Exact selection on a fixed grid)**.** Because $\mathcal{P}$ is a fixed finite grid, there is no intermediate asymptotic regime for $|\widehat{\phi} - \phi^*|$: either $\widehat{\phi} = \phi^*$, or the two values differ by at least the minimum spacing between distinct grid points, which is a fixed positive constant. Thus, the uncertainty in kernel selection is entirely captured by the probability of the exact-selection event in Lemma 14. On that event, $D_{i,j}(\widehat{\phi}) - D_{i,j}(\phi^*) = 0$ exactly. A nonzero discrepancy that vanishes with sample size would instead require a grid that becomes increasingly dense, together with identification and local curvature conditions on a continuum parameter space. We do not consider that regime, since the procedure in Section 3 searches over a fixed grid.

Lemma 15 (Hazard estimation rate)**.** Under Assumptions 4 and 5, for each fixed $\phi \in \mathcal{P}$ and each $\delta \in (0,1)$, there exist $C_\delta < \infty$ and $n_2(\delta)$ such that, with $r_{n,\delta} = C_\delta \sqrt{\frac{p_H}{n_{\mathrm{tr}}}}$, $\mathbb{P}\left(\|\widehat{\theta}(\phi) - \theta^*(\phi)\|_2 \le r_{n,\delta}\right) \ge 1 - \delta$ for all $n_{\mathrm{tr}} \ge n_2(\delta)$.

Proof. Fix $\phi \in \mathcal{P}$ and write $\theta^* = \theta^*(\phi)$, $\widehat{\theta} = \widehat{\theta}(\phi)$, $D_{i,j} = D_{i,j}(\phi)$,$\Delta = \widehat{\theta} - \theta^*$. Write $\widehat{L}_n(\theta) = \widehat{\ell}_n(\theta) + \lambda_n \omega(\theta)$ and $\widehat{\ell}_n(\theta) = \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathcal{I}_{\mathrm{fit}}} m_\theta(\mathcal{T}_i)$, with population counterpart $\ell(\theta) = \mathbb{E}[m_\theta(\mathcal{T}_i)]$. For sampling unit i,

$$m_\theta(\mathcal{T}_i) = -\sum_{j \in \mathcal{J}_i} \left[Z_{i,j} \log \lambda_\theta(i,j) + (1 - Z_{i,j}) \log\{1 - \lambda_\theta(i,j)\}\right].$$

By Assumption 4, the unit-level contributions $m_\theta(\mathcal{T}_i)$ are i.i.d. conditional on the fixed upstream feature pipeline, and each unit contains at most $N_{\max}$ at-risk observations. We first show the empirical strong convexity. For each sampling unit,

$$\nabla^2 m_\theta(\mathcal{T}_i) = \sum_{j \in \mathcal{J}_i} \sigma'\left(\theta^\top D_{i,j}\right) D_{i,j} D_{i,j}^\top \succeq 0,$$

so $\widehat{\ell}_n$ is convex. By Lemma 13 and the compactness of Θ, $|\theta^\top D_{i,j}| \le M_\theta M_D$ uniformly over $\theta \in \Theta$. Hence, $\sigma'(\theta^\top D_{i,j}) \ge c_\sigma := \sigma'(M_\theta M_D) > 0$, and therefore $\nabla^2 \widehat{\ell}_n(\theta) \succeq c_\sigma \widehat{\Sigma}_n$ and $\widehat{\Sigma}_n = \frac{1}{n_{\mathrm{tr}}} \sum_i \sum_{j \in \mathcal{J}_i} D_{i,j} D_{i,j}^\top$, for every $\theta \in \Theta$. The unit-level matrices $\sum_{j \in \mathcal{J}_i} D_{i,j} D_{i,j}^\top$ are i.i.d. positive semidefinite random matrices with operator norm bounded by $N_{\max} M_D^2$. Since p_H is fixed,

$\|\widehat{\Sigma}_n - \Sigma(\phi)\|_{\mathrm{op}} = O_{\mathbb{P}}(n_{\mathrm{tr}}^{-1/2})$. Together with $\lambda_{\min}\{\Sigma(\phi)\} \geq c_D$, this implies that, for sufficiently large n_{tr}, with probability at least $1 - \delta/2$, $\nabla^2 \widehat{\ell}_n(\theta) \succeq \mu I$ for all $\theta \in \Theta$, where $\mu := \frac{1}{2} c_\sigma c_D > 0$. Let $\mathcal{G}_n$ denote this event. On $\mathcal{G}_n$, $\widehat{\ell}_n$ is μ-strongly convex on Θ. Then, derive a basic inequality.

Since $\widehat{\theta}$ minimizes $\widehat{L}_n$ over Θ and $\theta^* \in \Theta$,

$$\widehat{\ell}_n(\widehat{\theta}) - \widehat{\ell}_n(\theta^*) \leq \lambda_n \left\{ \omega(\theta^*) - \omega(\widehat{\theta}) \right\} \leq \lambda_n L_\omega \|\Delta\|_2.$$

On $\mathcal{G}_n$, strong convexity gives $\widehat{\ell}_n(\widehat{\theta}) - \widehat{\ell}_n(\theta^*) \geq \nabla \widehat{\ell}_n(\theta^*)^\top \Delta + \frac{\mu}{2} \|\Delta\|_2^2$. Combining the two inequalities and applying Cauchy–Schwarz yields $\frac{\mu}{2} \|\Delta\|_2^2 \leq \left\{ \|\nabla \widehat{\ell}_n(\theta^*)\|_2 + \lambda_n L_\omega \right\} \|\Delta\|_2$. Therefore,

$$\|\Delta\|_2 \leq \frac{2}{\mu} \{ \|\nabla \widehat{\ell}_n(\theta^*)\|_2 + \lambda_n L_\omega \}. \tag{20}$$

Then, we derive the score concentration. Because θ^* is an interior minimizer of ℓ, $\nabla \ell(\theta^*) = 0$.

Thus, $\nabla \widehat{\ell}_n(\theta^*)$ is an average of centered i.i.d. unit-level scores,

$$\nabla m_{\theta^*}(\mathcal{T}_i) = - \sum_{j \in \mathcal{J}_i} \{ Z_{i,j} - \lambda_{\theta^*}(i,j) \} D_{i,j},$$

each bounded in norm by $N_{\max} M_D$. Hence $\mathbb{E} \left\| \nabla \widehat{\ell}_n(\theta^*) \right\|_2^2 = O\left(\frac{p_H}{n_{\mathrm{tr}}} \right)$. Therefore, for every $\delta \in (0,1)$, there exists $C'_\delta < \infty$ such that

$$\mathbb{P}\left(\|\nabla \widehat{\ell}_n(\theta^*)\|_2 \leq C'_\delta \sqrt{\frac{p_H}{n_{\mathrm{tr}}}} \right) \geq 1 - \frac{\delta}{2}$$

for all sufficiently large n_{tr}. Intersecting the score-concentration event with $\mathcal{G}_n$ and applying the union bound gives an event of probability at least $1 - \delta$. On this event, equation 20 and $\lambda_n = O(n_{\mathrm{tr}}^{-1/2})$ imply $\|\widehat{\theta}(\phi) - \theta^*(\phi)\|_2 \leq C_\delta \sqrt{\frac{p_H}{n_{\mathrm{tr}}}}$ for a finite constant C_δ. This proves the claim. □

C Full Empirical Results

Table A1: Saturation and marginal separation of the current-turn uncertainty feature $\widehat{U}$ for Qwen in the HalluHard landmark setting ($t \geq 3$). Each at-risk state draws $K = 5$ diagnostic samples, so $\widehat{U}$ attains at most $\log 5 + 4/10 \approx 2.009$, reached when all five samples fall in distinct semantic classes; the saturation rate is the fraction of at-risk states at that value. The last two columns report mean $\widehat{U}$ over first-failure and non-failure turns.

Domain	Saturation rate	$\overline{U}$ (failure)	$\overline{U}$ (non-failure)
Legal	10.9%	0.9583	0.7452
Medical	13.3%	1.0949	0.7803
Coding	38.5%	1.4572	1.3562
Research	48.3%	1.4879	1.4920

C.1 Experimental Setting

Additional Experimental Details. For HalluHard, since the conversation length T_i is fixed by protocol, censoring is administrative, so conditioning in equation 1 induces no selection. Diagnostic responses are sampled using the same system prompt, chat template, and decoding parameters as the released response, which is excluded from semantic clustering. For PRM800K, neutral (0) and positive (+1) steps are treated as failure-free. Exact matching after text normalization provides a stricter equivalence criterion than pairwise entailment clustering. We assess robustness across

minimum reference-pool thresholds $n_{\min} \in \{2, 3\}$ and both capped and unrestricted sibling-pool regimes; the incremental benefit of H persists across these settings within our discrete-time hazard model. Kernel parameters are selected on the training set from a prespecified grid in $(0, 1)$. Reported results are means over 20 random splits, with standard deviations in parentheses.

Additional Evaluation Details. At the population level, MWR corresponds to $\Pr(\Delta = 1, \tau^{\text{warn}} > \tau)$, so the target level α controls a marginal missed-first-failure probability rather than a miss probability conditional on failure. In the empirical comparisons, a construction is deemed admissible at level α only if its mean MWR across splits does not exceed α; constructions that reduce warning burden by missing additional first failures are excluded from the admissible comparison. The warning rate (WR) is the fraction of test units receiving a warning within the observed horizon, while the unnecessary-warning rate (UWR) is the corresponding fraction restricted to units that never fail. Higher AUROC and AUPR indicate better first-failure ranking, whereas lower WR and UWR indicate lower warning burden. The raw uncertainty baselines U, H, and their rank-normalized average are used directly as risk scores without a hazard layer. Within the hazard-model comparisons, $\widehat{\lambda}\{f + U\}$ and $\widehat{\lambda}\{f + U + H\}$ are the fitted counterparts of the propagation-oblivious class $\mathcal{S}_0$ and propagation-aware class $\mathcal{S}_H$, respectively, in Definition 1.

For the reproducibility of the factuality labelling, we re-judge 30 randomly sampled claims per cell with the same judge. Each claim c receives an original label A_c and a repeated label A'_c, with $A_c, A'_c \in \{\text{correct}, \text{incorrect}, \text{uncertain}\}$. Three-way agreement is $m^{-1} \sum_c \mathbf{1}\{A_c = A'_c\}$ with $m = 30$. Collapsing to a binary factual-error label $B_c = \mathbf{1}\{A_c = \text{incorrect}\}$ gives binary agreement $m^{-1} \sum_c \mathbf{1}\{B_c = B'_c\}$. Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$ (Cohen, 1960) reports agreement beyond the level induced by the marginal label frequencies. Repeat-judgement disagreements concentrate on claims originally labelled *uncertain*, which the binary collapse maps to "not an error" under both readings, so they do not propagate to $E_{i,t}$.

For the uncertainty features, two diagnostic responses are placed in the same class when they are mutually entailing at the claim level, following the bidirectional-entailment construction of Kuhn et al. (2023); entailment is evaluated by the same judge model with a prompt that asks whether two responses give the same factual answer, not whether either is correct. NLI features were constructed for all at-risk states. In the chain-of-thought setting, a pool of n chains costs $\binom{n}{2}$ pairwise judgments. Summing over the 98,576 (problem, step) pools of the $K_p \geq 5$ setting (n: median 3, mean 6.1, max 469) gives 2.34×10^7, against 65,760 for the eight multi-turn cells combined, where $K = 5$. We therefore substitute exact match after text normalisation. The substitution is conservative in direction, since exact match can only split classes an entailment judge would merge, so $\widehat{U}$ and $\widehat{q}$ are upward-biased. It is, however, coarse: on 55% of at-risk steps every reference falls in its own class, so $\widehat{U}$ and $\widehat{q}$ attain their maxima and reduce to functions of the pool size. Since $n_{p,k,t}$ decreases with t, these saturated states are concentrated at later steps.

Let $\tau_i^{\text{warn}} = \inf\{t : R_{i,t} > \widehat{\eta}\}$ denote the first warning time under the calibrated threshold $\widehat{\eta}$, with $\inf \varnothing = \infty$. The three decision metrics of main text are

$$\begin{aligned} \text{MWR} &= \frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbf{1}\{\Delta_i = 1, \tau_i^{\text{warn}} > \tau_i\}, \\ \text{WR} &= \frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbf{1}\{\tau_i^{\text{warn}} \leq Y_i\}, \\ \text{UWR} &= \frac{\sum_{i \in \mathcal{I}_{\text{test}}} (1 - \Delta_i) \mathbf{1}\{\tau_i^{\text{warn}} \leq Y_i\}}{\sum_{i \in \mathcal{I}_{\text{test}}} (1 - \Delta_i)}. \end{aligned} \tag{21}$$

We additionally report the mean lead time $\text{Lead} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\tau_i - \tau_i^{\text{warn}})$, where $\mathcal{S} = \{i : \Delta_i = 1, \tau_i^{\text{warn}} \leq \tau_i\}$, averaged over failures that are successfully warned. Lead is descriptive and is not used to rank constructions: a value of zero means the warning is issued immediately before the failing response, and larger values mean earlier intervention. We report it so that a low warning rate can be distinguished from a rule that warns only at the last possible step.

Full numerical results are reported in the following sections. We additionally report the non-modal semantic mass $q_{i,t} = 1 - \max_c \widehat{p}_{i,t,c}$, the mass outside the modal semantic class, which is the quantity propagated by H.

Table A2 reports agreement by cell. Three-way agreement ranges from 0.733 to 0.880 and binary factual-error agreement from 0.767 to 0.967; the corresponding κ values are 0.60–0.82 and 0.46–0.93. Agreement is lowest in the Llama coding cell, which also has the most administrative censoring.

Table A2: Judge self-consistency by cell. Claims are re-judged with the same judge, with sibling claims of the same response in context. Three-way agreement is over {`correct`, `incorrect`, `uncertain`}; the binary label collapses to $\mathbf{1}\{$`incorrect`$\}$.

Model	Domain	3-way	κ_3	Binary	κ_{bin}
Qwen	Medical	0.800	0.700	0.867	0.714
	Legal	0.867	0.800	0.900	0.781
	Research	0.880	0.816	0.960	0.896
	Coding	0.833	0.750	0.833	0.651
Llama	Medical	0.767	0.650	0.833	0.595
	Legal	0.833	0.750	0.967	0.927
	Research	0.800	0.700	0.933	0.857
	Coding	0.733	0.600	0.767	0.462

Event prevalence and the comparability of AUPR. AUPR is bounded below by the event prevalence of the fold on which it is computed, so it is only comparable between settings that share that prevalence. Averaged over the twenty test folds, the random-ranker AUPR on the full risk set is 0.1455 (Qwen) against 0.3165 (Llama) for medical, 0.2112 against 0.5790 for legal, 0.2524 against 0.5806 for coding, and 0.0769 against 0.2858 for research; on the $t \geq 3$ landmark populations the corresponding values are 0.1260 against 0.2310, 0.1711 against 0.3835, 0.1935 against 0.5079, and 0.0845 against 0.2231. Llama fails between two and three times as often as Qwen in every domain, so its higher AUPR reflects a denser event stream, and each column should be read against its own baseline. The same caution applies across domains, which span a range of roughly sevenfold from research to coding. Within a model the landmark restriction lowers prevalence everywhere except research, where it rises slightly for Qwen ($0.0769 \rightarrow 0.0845$), consistent with research failures arriving later in the conversation.

Table A3: Event prevalence in the held-out folds of the turn-level prediction experiments. π is the mean over 20 random conversation-level splits of the fraction of at-risk test turns carrying the first-failure event, with the across-split standard deviation in parentheses; it is the AUPR attained by a random ranker and therefore the floor against which each AUPR column must be read. Counts of test turns and events are also means over the 20 splits.

Domain	Model	Population	Split	Test turns	Events	π
Medical	Qwen	full ($t \geq 1$)(250)	150/50/50	199.7	29.0	0.1455(0.0064)
	Qwen	landmark ($t \geq 3$)(176)	106/35/35	111.2	14.0	0.1260(0.0039)
	Llama	full ($t \geq 1$)(249)	149/50/50	133.3	42.0	0.3165(0.0225)
	Llama	landmark ($t \geq 3$)(106)	64/21/21	56.5	13.0	0.2310(0.0132)
Legal	Qwen	full ($t \geq 1$)(250)	150/50/50	170.8	36.0	0.2112(0.0086)
	Qwen	landmark ($t \geq 3$)(144)	86/29/29	87.9	15.0	0.1711(0.0094)
	Llama	full ($t \geq 1$)(249)	149/50/50	84.3	48.4	0.5790(0.0571)
	Llama	landmark ($t \geq 3$)(40)	24/8/8	17.1	6.4	0.3835(0.0680)
Coding	Qwen	full ($t \geq 1$)(299)	179/60/60	186.6	47.0	0.2524(0.0120)
	Qwen	landmark ($t \geq 3$)(149)	89/30/30	88.1	17.0	0.1935(0.0104)
	Llama	full ($t \geq 1$)(296)	178/59/59	98.8	57.0	0.5806(0.0481)
	Llama	landmark ($t \geq 3$)(49)	29/10/10	17.9	9.0	0.5079(0.0525)
Research	Qwen	full ($t \geq 1$)(250)	150/50/50	247.4	19.0	0.0769(0.0023)
	Qwen	landmark ($t \geq 3$)(219)	131/44/44	153.9	13.0	0.0845(0.0013)
	Llama	full ($t \geq 1$)(250)	150/50/50	145.3	41.4	0.2858(0.0179)
	Llama	landmark ($t \geq 3$)(114)	68/23/23	64.7	14.4	0.2231(0.0148)

Fold size differs as sharply and drives the reported standard deviations. In research and medical the landmark retains most of the corpus, 219 of 250 and 176 of 250 Qwen conversations, so a test fold carries 154 and 111 at-risk turns and the prevalence itself is stable across splits (across-split sd 0.0013 and 0.0039). In legal and coding the landmark is far more restrictive for Llama, where only 40 of 249 and 49 of 296 conversations survive to turn 3, leaving test folds of 17 and 18 at-risk turns containing 6.4 and 9.0 events on average. There the baseline itself moves from split to split (across-split sd 0.0680 and 0.0525), so part of the spread in those AUPR columns is variation in what a random ranker would have achieved rather than variation in the ranking.

C.2 Additional Experimental Results

In this section, we address the following question: How robust is the incremental predictive value of H across survival models and domains?

In this section, we are trying to answer the following question. How robust is the incremental predictive value of H across survival models and domains? Having shown in Q2 that the learned

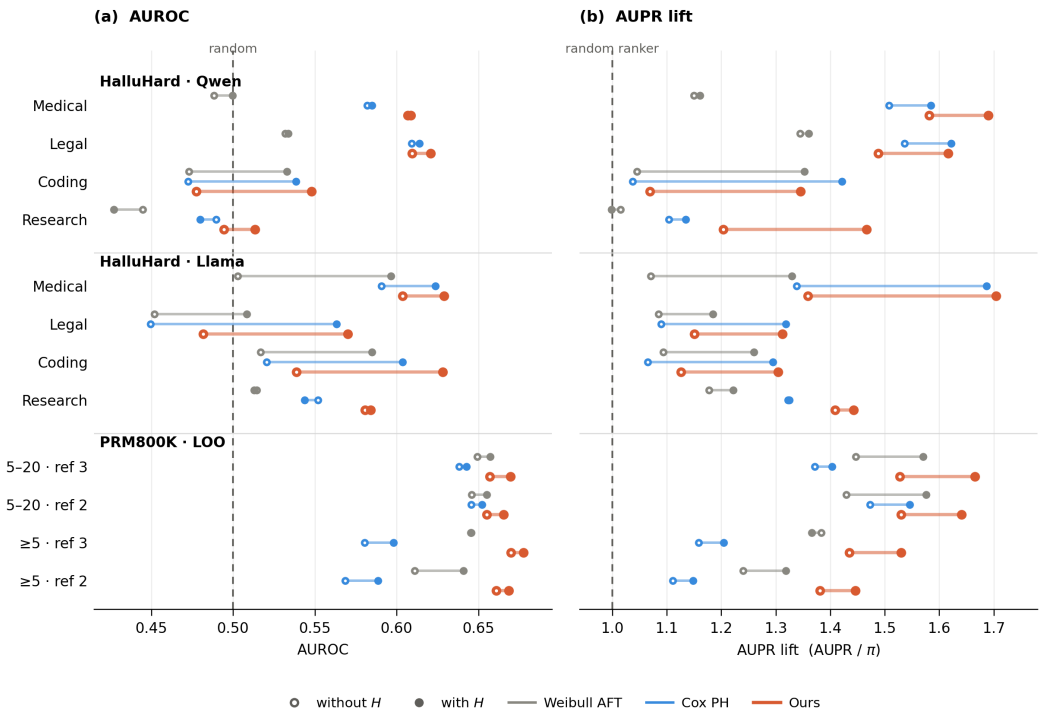

Figure 6: An ablation of the historical feature H across hazard families and settings.

hazard layer can extract useful predictive signal from the uncertainty features, we next address Q3 by examining how robust the incremental predictive value of the propagated feature H is across different survival models and domains. Figure 6 reports the effect of adding H beyond the current-turn uncertainty U on first-failure discrimination. For AUROC, the benefit of H exhibits a clear domain-dependent pattern. In Coding, adding H improves discrimination for all three survival models under both Qwen and Llama, making it the domain with the most consistently pronounced gains. In Legal, the improvement is particularly strong for Llama compared with Qwen, indicating that H is particularly beneficial for Llama in this setting. In Medical, all three survival models also benefit from H, with Weibull AFT showing the largest gain among the three models for both target LLMs. However, in the Research domain, adding H decreases AUROC for both Weibull AFT and Cox PH under Qwen, with smaller decreases for the same two models under Llama, while our discrete-time hazard model still improves in both cases. One possible explanation is that Research contains more open-ended steps, where semantically diverse continuations may all remain correct, so historical instability may partly reflect benign diversity. In the unrestricted PRM800K setting with

minimum reference pool 3, the Weibull AUROC is essentially unchanged, while AUPR decreases slightly. A contributing factor may be that the leave-one-out reference pool shrinks with step depth, so that H at deeper steps is estimated from fewer sibling chains and may carry greater measurement noise. Moreover, AUPR lift exhibits a broadly similar pattern. In Coding, adding H consistently improves performance across all three survival models and both target LLMs, with the gains again more pronounced for Llama. Legal and Medical show a similar model dependent pattern, where the contribution of H is generally stronger for Llama than for Qwen. Research remains the main exception, with smaller and less consistent changes across the alternative survival models. A similar pattern also appears in PRM800K, where H is generally beneficial across the evaluated settings, with the main exception being the weaker Weibull AFT result in the unrestricted cohort discussed above.

C.3 Results on Conversations Surviving to Turn 3

Table A4: Turn-level prediction on the landmark population $t \geq 3$ for the medical domain, with pre-landmark turns dropped. For Qwen, 176 conversations survive to turn 3, using 106/35/35 train/validation/test conversation-level splits; values are means over 20 random splits. For Llama, 106 conversations survive to turn 3, using 64/21/21 splits; values are means over 20 random splits. Across-split standard deviations are reported in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.5592(0.0749)	0.1624(0.0401)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5951(0.0804)	0.1673(0.0353)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.6069(0.0855)	0.1993(0.0610)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.6072(0.0960)	0.2070(0.0554)
	$R_{i,t} = q_{i,t}$	0.5565(0.0738)	0.1562(0.0334)
	$R_{i,t} = H_{i,t}$	0.5957(0.0816)	0.1826(0.0390)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5964(0.0821)	0.1825(0.0452)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6086(0.0968)**	**0.2129(0.0625)**
Llama	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.4428(0.0695)	0.2164(0.0196)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5675(0.0740)	0.2831(0.0609)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.6036(0.0903)	0.3140(0.0821)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.6235(0.0767)	0.3890(0.1161)
	$R_{i,t} = q_{i,t}$	0.4433(0.0689)	0.2191(0.0226)
	$R_{i,t} = H_{i,t}$	0.3689(0.0683)	0.1988(0.0194)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.3950(0.0725)	0.2011(0.0220)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6290(0.0820)**	**0.3936(0.1191)**

Table A5: Turn-level prediction across hazard-model benchmarks on the landmark population $t \geq 3$ for the medical domain. For Qwen, 176 conversations survive to turn 3, using 106/35/35 train/validation/test conversation-level splits; values are means over 20 random splits. For Llama, 106 conversations survive to turn 3, using 64/21/21 splits; values are means over 20 random splits. Across-split standard deviations are reported in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.4883(0.0800)	0.1449(0.0371)
		Cox PH	0.5819(0.0862)	0.1900(0.0559)
		Ours	**0.6069(0.0855)**	**0.1993(0.0610)**
	with H	Weibull AFT	0.4996(0.0852)	0.1463(0.0437)
		Cox PH	0.5849(0.0825)	0.1997(0.0579)
		Ours	**0.6086(0.0968)**	**0.2129(0.0625)**
Llama	without H	Weibull AFT	0.5028(0.0842)	0.2474(0.0387)
		Cox PH	0.5907(0.0966)	0.3091(0.0829)
		Ours	**0.6036(0.0903)**	**0.3140(0.0821)**
	with H	Weibull AFT	0.5966(0.0683)	0.3071(0.0486)
		Cox PH	0.6237(0.0833)	0.3897(0.1166)
		Ours	**0.6290(0.0820)**	**0.3936(0.1191)**

Table A6: Turn-level prediction on the landmark population $t \geq 3$, with pre-landmark turns dropped. For Qwen, 144 conversations survive to turn 3, using 86/29/29 splits. For Llama, 40 conversations survive to turn 3, using 24/8/8 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.6104(0.0886)	0.2478(0.0520)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5302(0.0638)	0.1879(0.0253)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.6094(0.0898)	0.2546(0.0611)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5938(0.0677)	0.2302(0.0430)
	$R_{i,t} = q_{i,t}$	0.5751(0.0760)	0.2186(0.0400)
	$R_{i,t} = H_{i,t}$	0.5760(0.0811)	0.2427(0.0671)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5899(0.0826)	0.2672(0.0755)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6209(0.0841)**	**0.2765(0.0771)**
Llama	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.4907(0.1387)	0.4359(0.1065)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5124(0.0751)	0.4207(0.1058)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.4818(0.0916)	0.4413(0.1434)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5495(0.1204)	0.4961(0.1536)
	$R_{i,t} = q_{i,t}$	0.4948(0.1340)	0.4375(0.0947)
	$R_{i,t} = H_{i,t}$	0.4874(0.1868)	0.4615(0.1701)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4649(0.1271)	0.4339(0.1182)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.5700(0.1425)**	**0.5033(0.1738)**

Table A7: Turn-level prediction across hazard-model benchmarks on the landmark population $t \geq 3$. For Qwen, 144 conversations survive to turn 3, using 86/29/29 splits; for Llama, 40 conversations survive to turn 3, using 24/8/8 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.5318(0.0755)	0.2301(0.0483)
		Cox PH	0.6092(0.0838)	**0.2628(0.0554)**
		Ours	**0.6094(0.0898)**	0.2546(0.0611)
	with H	Weibull AFT	0.5338(0.0751)	0.2328(0.0491)
		Cox PH	0.6139(0.0795)	**0.2775(0.0708)**
		Ours	**0.6209(0.0841)**	0.2765(0.0771)
Llama	without H	Weibull AFT	0.4518(0.1317)	0.4162(0.1216)
		Cox PH	0.4494(0.0802)	0.4179(0.0969)
		Ours	**0.4818(0.0916)**	**0.4413(0.1434)**
	with H	Weibull AFT	0.5084(0.1400)	0.4544(0.1092)
		Cox PH	0.5632(0.1417)	**0.5056(0.1806)**
		Ours	**0.5700(0.1425)**	0.5033(0.1738)

Table A8: Turn-level prediction on the landmark population $t \geq 3$ for the coding domain, with pre-landmark turns dropped. For Qwen, 149 conversations survive to turn 3, using 89/30/30 splits. For Llama, 49 conversations survive to turn 3, using 29/10/10 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.5310(0.0634)	0.2213(0.0400)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.4526(0.0637)	0.1863(0.0276)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.4774(0.0663)	0.2068(0.0384)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5315(0.0909)	0.2591(0.0611)
	$R_{i,t} = q_{i,t}$	0.5388(0.0603)	0.2230(0.0379)
	$R_{i,t} = H_{i,t}$	0.4177(0.0713)	0.1774(0.0309)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4856(0.0670)	0.2048(0.0348)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.5479(0.0901)**	**0.2603(0.0814)**
Llama	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.4426(0.1041)	0.5036(0.0703)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5680(0.0959)	0.5598(0.0915)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.5388(0.1205)	0.5720(0.1125)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.6150(0.1348)	0.6517(0.1456)
	$R_{i,t} = q_{i,t}$	0.4387(0.1019)	0.5006(0.0711)
	$R_{i,t} = H_{i,t}$	0.3508(0.1061)	0.4665(0.0718)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.3424(0.1027)	0.4916(0.0962)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6281(0.0951)**	**0.6626(0.1048)**

Table A9: Turn-level prediction across hazard-model benchmarks on the landmark population $t \geq 3$ for the coding domain. For Qwen, 149 conversations survive to turn 3, using 89/30/30 splits; for Llama, 49 conversations survive to turn 3, using 29/10/10 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.4731(0.0586)	0.2023(0.0239)
		Cox PH	0.4725(0.0546)	0.2007(0.0309)
		Ours	**0.4774(0.0663)**	**0.2068(0.0384)**
	with H	Weibull AFT	0.5330(0.0854)	**0.2618(0.0719)**
		Cox PH	0.5384(0.0649)	0.2550(0.0674)
		Ours	**0.5479(0.0901)**	0.2603(0.0814)
Llama	without H	Weibull AFT	0.5167(0.1153)	0.5555(0.1058)
		Cox PH	0.5206(0.0882)	0.5412(0.0947)
		Ours	**0.5388(0.1205)**	**0.5720(0.1125)**
	with H	Weibull AFT	0.5848(0.1326)	0.6399(0.1209)
		Cox PH	0.6035(0.0971)	0.6575(0.1111)
		Ours	**0.6281(0.0951)**	**0.6626(0.1048)**

Table A10: Turn-level prediction on the landmark population $t \geq 3$ for the research domain, with pre-landmark turns dropped. For Qwen, 219 conversations survive to turn 3, using 131/44/44 splits. For Llama, 114 conversations survive to turn 3 (317 at-risk turns), using 68/23/23 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\text{nli,mm},i,t}$	0.4601(0.0677)	0.0821(0.0111)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.4925(0.0716)	0.0885(0.0145)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t}\}$	0.4944(0.0882)	0.1017(0.0269)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\text{nli},i,t}\}$	**0.5296(0.0555)**	0.1153(0.0334)
	$R_{i,t} = q_{i,t}$	0.4494(0.0647)	0.0802(0.0116)
	$R_{i,t} = H_{i,t}$	0.4329(0.0730)	0.0782(0.0121)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4270(0.0557)	0.0781(0.0120)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t} + H_{\text{nli},i,t}\}$	0.5133(0.0633)	**0.1239(0.0485)**
Llama	$R_{i,t} = U_{\text{nli,mm},i,t}$	0.5737(0.0614)	0.2872(0.0515)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.4804(0.0873)	0.2289(0.0396)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t}\}$	0.5806(0.0688)	0.3143(0.0592)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\text{nli},i,t}\}$	0.5359(0.0899)	0.2899(0.0644)
	$R_{i,t} = q_{i,t}$	0.5769(0.0627)	0.2928(0.0547)
	$R_{i,t} = H_{i,t}$	0.5452(0.0612)	0.2975(0.0454)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5697(0.0695)	0.3120(0.0492)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t} + H_{\text{nli},i,t}\}$	**0.5841(0.0864)**	**0.3219(0.0805)**

Table A11: Turn-level prediction across hazard-model benchmarks on the landmark population $t \geq 3$ for the research domain. For Qwen, 219 conversations survive to turn 3, using 131/44/44 splits; for Llama, 114 conversations survive to turn 3, using 68/23/23 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.4447(0.0555)	0.0858(0.0133)
		Cox PH	0.4895(0.0802)	0.0933(0.0197)
		Ours	**0.4944(0.0882)**	**0.1017(0.0269)**
	with H	Weibull AFT	0.4271(0.0624)	0.0844(0.0135)
		Cox PH	0.4798(0.0761)	0.0959(0.0260)
		Ours	**0.5133(0.0633)**	**0.1239(0.0485)**
Llama	without H	Weibull AFT	0.5144(0.0633)	0.2628(0.0446)
		Cox PH	0.5520(0.0862)	0.2955(0.0625)
		Ours	**0.5806(0.0688)**	**0.3143(0.0592)**
	with H	Weibull AFT	0.5128(0.0653)	0.2726(0.0514)
		Cox PH	0.5438(0.0887)	0.2950(0.0734)
		Ours	**0.5841(0.0864)**	**0.3219(0.0805)**

C.3.1 Conformal Results

Table A12: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H on the landmark risk set ($t \geq 3$) for the medical domain. Qwen results use 176 conversations and 20 random 106/35/35 conversation-level splits; Llama results use 106 conversations and 20 random 64/21/21 conversation-level splits. Each cell reports the across-split mean, with the standard deviation in parentheses.

LLM	Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
Qwen	0.10	with H	Weibull AFT	0.077(0.055)	0.885(0.105)	0.937(0.133)	0.965(0.301)
			Cox PH	0.076(0.054)	**0.754(0.135)**	**0.716(0.153)**	1.050(0.297)
			Ours	0.074(0.054)	0.755(0.131)	**0.716(0.147)**	1.037(0.293)
		without H	Weibull AFT	0.050(0.060)	0.937(0.090)	0.979(0.083)	0.967(0.281)
			Cox PH	0.026(0.055)	0.930(0.146)	0.927(0.161)	0.990(0.285)
			Ours	0.025(0.054)	**0.926(0.154)**	**0.917(0.172)**	0.981(0.276)
	0.20	with H	Weibull AFT	0.187(0.073)	0.706(0.161)	0.822(0.233)	0.784(0.368)
			Cox PH	0.173(0.067)	0.473(0.138)	0.410(0.141)	0.974(0.367)
			Ours	0.174(0.066)	**0.469(0.134)**	**0.406(0.138)**	0.967(0.366)
		without H	Weibull AFT	0.177(0.072)	0.742(0.158)	0.865(0.220)	0.736(0.314)
			Cox PH	0.153(0.082)	0.554(0.172)	0.512(0.181)	0.945(0.368)
			Ours	0.161(0.083)	**0.537(0.167)**	**0.496(0.169)**	0.921(0.383)
Llama	0.10	with H	Weibull AFT	0.083(0.079)	0.881(0.100)	0.906(0.090)	0.753(0.248)
			Cox PH	0.083(0.084)	0.826(0.138)	0.762(0.185)	0.766(0.246)
			Ours	0.086(0.084)	**0.821(0.132)**	**0.756(0.165)**	0.761(0.246)
		without H	Weibull AFT	0.062(0.080)	0.917(0.098)	0.944(0.111)	0.816(0.228)
			Cox PH	0.050(0.066)	0.867(0.088)	0.781(0.127)	0.790(0.252)
			Ours	0.052(0.065)	**0.857(0.086)**	**0.762(0.128)**	0.788(0.254)
	0.20	with H	Weibull AFT	0.140(0.088)	0.779(0.138)	0.787(0.177)	0.710(0.255)
			Cox PH	0.136(0.089)	0.717(0.152)	0.613(0.210)	0.711(0.250)
			Ours	0.138(0.089)	**0.712(0.158)**	**0.606(0.216)**	0.723(0.245)
		without H	Weibull AFT	0.148(0.094)	0.817(0.131)	0.906(0.140)	0.865(0.348)
			Cox PH	0.140(0.082)	0.750(0.108)	0.713(0.122)	0.849(0.362)
			Ours	0.138(0.085)	**0.748(0.108)**	**0.700(0.131)**	0.847(0.364)

Table A13: Conformal warning results on the landmark risk set ($t \geq 3$) for the medical domain. Qwen results use 176 conversations still at risk at turn 3 and 20 random 106/35/35 conversation-level splits; Llama results use 106 conversations still at risk at turn 3 and 20 random 64/21/21 conversation-level splits. Each cell reports the across-split mean, with the standard deviation in parentheses.

LLM	Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
Qwen	0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.045(0.065)	0.897(0.142)	0.904(0.137)	0.983(0.279)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.996(0.266)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.025(0.054)	0.926(0.154)	0.917(0.172)	0.981(0.276)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.071(0.060)	0.786(0.136)	0.761(0.147)	1.020(0.309)
		$R_{i,t} = q_{i,t}$	0.042(0.061)	0.903(0.129)	0.910(0.123)	0.979(0.278)
		$R_{i,t} = H_{i,t}$	0.073(0.055)	0.804(0.114)	0.795(0.120)	1.004(0.304)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.081(0.054)	0.789(0.123)	0.745(0.138)	1.016(0.310)
		Our Model	0.074(0.054)	**0.755(0.131)**	**0.716(0.147)**	1.037(0.293)
	0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.165(0.070)	0.563(0.156)	0.547(0.169)	0.778(0.344)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.996(0.266)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.161(0.083)	0.537(0.167)	0.496(0.169)	0.921(0.383)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.180(0.071)	0.485(0.151)	0.441(0.160)	1.084(0.436)
		$R_{i,t} = q_{i,t}$	0.150(0.065)	0.618(0.147)	0.613(0.162)	0.814(0.326)
		$R_{i,t} = H_{i,t}$	0.182(0.064)	0.509(0.141)	0.485(0.154)	1.036(0.392)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.186(0.069)	0.499(0.132)	0.464(0.133)	0.718(0.363)
		Our Model	0.174(0.066)	**0.469(0.134)**	**0.406(0.138)**	0.967(0.366)
Llama	0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.792(0.210)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.792(0.210)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.052(0.065)	0.857(0.086)	0.762(0.128)	0.788(0.254)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.071(0.082)	0.852(0.110)	0.800(0.124)	0.736(0.230)
		$R_{i,t} = q_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.792(0.210)
		$R_{i,t} = H_{i,t}$	0.024(0.061)	0.969(0.079)	0.981(0.061)	0.786(0.255)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.050(0.072)	0.933(0.105)	0.956(0.124)	0.792(0.261)
		Our Model	0.086(0.084)	**0.821(0.132)**	**0.756(0.165)**	0.761(0.246)
	0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.062(0.098)	0.921(0.127)	0.956(0.102)	0.752(0.229)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	0.792(0.210)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.138(0.085)	0.748(0.108)	0.700(0.131)	0.847(0.364)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.136(0.096)	0.721(0.173)	0.625(0.229)	0.721(0.240)
		$R_{i,t} = q_{i,t}$	0.062(0.098)	0.921(0.127)	0.956(0.102)	0.752(0.229)
		$R_{i,t} = H_{i,t}$	0.131(0.107)	0.848(0.119)	0.944(0.076)	0.845(0.312)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.181(0.121)	0.769(0.159)	0.869(0.131)	0.797(0.297)
		Our Model	0.138(0.089)	**0.712(0.158)**	**0.606(0.216)**	0.723(0.245)

Our Model denotes $R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$.

Table A14: Conformal warning results on the landmark risk set ($t \geq 3$) for the legal domain. Results are based on the 144 Qwen conversations that are still at risk at turn 3. Each cell reports the mean over 20 random 86/29/29 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.048(0.073)	0.931(0.103)	0.957(0.067)	1.107(0.300)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.197(0.263)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.050(0.079)	0.914(0.130)	0.925(0.132)	1.126(0.261)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.086(0.088)	0.872(0.149)	0.914(0.142)	1.231(0.267)
	$R_{i,t} = q_{i,t}$	0.048(0.073)	0.931(0.103)	0.957(0.067)	1.114(0.296)
	$R_{i,t} = H_{i,t}$	0.071(0.064)	0.900(0.097)	0.939(0.091)	1.189(0.255)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.074(0.061)	0.884(0.092)	0.914(0.075)	1.086(0.302)
	Our Model	0.093(0.073)	**0.838(0.111)**	**0.857(0.111)**	1.097(0.303)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.152(0.090)	0.743(0.152)	0.782(0.151)	0.934(0.332)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.197(0.263)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.159(0.106)	0.697(0.160)	0.700(0.151)	1.010(0.274)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.153(0.098)	0.721(0.154)	0.739(0.140)	1.110(0.293)
	$R_{i,t} = q_{i,t}$	0.141(0.083)	0.778(0.124)	0.832(0.121)	0.976(0.342)
	$R_{i,t} = H_{i,t}$	0.159(0.096)	0.712(0.168)	0.732(0.173)	1.008(0.278)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.176(0.083)	0.667(0.122)	0.675(0.126)	0.986(0.382)
	Our Model	0.186(0.090)	**0.650(0.147)**	**0.661(0.159)**	1.064(0.333)

Table A15: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on the landmark risk set ($t \geq 3$) for the legal domain (144 conversations still at risk at turn 3). Each cell is the mean over 20 random 86/29/29 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Historical feature	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.076(0.067)	0.916(0.085)	0.982(0.056)	1.017(0.322)
		Cox PH	0.084(0.075)	0.869(0.119)	0.904(0.114)	1.111(0.299)
		Ours	0.093(0.073)	**0.838(0.111)**	**0.857(0.111)**	1.097(0.303)
	without H	Weibull AFT	0.060(0.062)	0.931(0.083)	0.982(0.056)	1.044(0.323)
		Cox PH	0.038(0.047)	0.936(0.093)	0.946(0.118)	1.138(0.250)
		Ours	0.050(0.079)	**0.914(0.130)**	**0.925(0.132)**	1.126(0.261)
0.20	with H	Weibull AFT	0.172(0.097)	0.798(0.131)	0.939(0.117)	0.795(0.321)
		Cox PH	0.179(0.093)	0.666(0.141)	0.679(0.147)	1.048(0.338)
		Ours	0.186(0.090)	**0.650(0.147)**	**0.661(0.159)**	1.064(0.333)
	without H	Weibull AFT	0.155(0.099)	0.819(0.131)	0.946(0.111)	0.829(0.319)
		Cox PH	0.166(0.106)	0.703(0.165)	0.729(0.165)	1.016(0.255)
		Ours	0.159(0.106)	**0.697(0.160)**	**0.700(0.151)**	1.010(0.274)

Table A16: Conformal warning results on the landmark risk set ($t \geq 3$) for the coding domain. Results are based on the 149 Qwen conversations that are still at risk at turn 3. Each cell reports the mean over 20 random 89/30/30 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.067(0.065)	0.928(0.073)	0.988(0.038)	1.095(0.260)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.141(0.226)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.067(0.082)	0.920(0.102)	0.969(0.080)	1.155(0.298)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.047(0.057)	0.912(0.094)	0.904(0.129)	1.112(0.256)
	$R_{i,t} = q_{i,t}$	0.067(0.065)	0.928(0.073)	0.988(0.038)	1.095(0.260)
	$R_{i,t} = H_{i,t}$	0.087(0.067)	0.893(0.086)	0.954(0.052)	1.130(0.247)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.095(0.069)	0.893(0.081)	0.973(0.063)	1.122(0.283)
	Our Model	0.073(0.067)	**0.878(0.112)**	**0.888(0.118)**	1.004(0.237)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.162(0.081)	0.797(0.105)	0.904(0.086)	1.060(0.304)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.032(0.098)	0.968(0.098)	1.000(0.000)	1.154(0.231)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.162(0.089)	0.788(0.108)	0.885(0.140)	1.202(0.283)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.145(0.086)	0.742(0.143)	0.738(0.182)	0.949(0.248)
	$R_{i,t} = q_{i,t}$	0.142(0.070)	0.825(0.078)	0.923(0.061)	1.060(0.304)
	$R_{i,t} = H_{i,t}$	0.185(0.073)	0.740(0.110)	0.827(0.122)	1.151(0.265)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.180(0.085)	0.742(0.136)	0.819(0.146)	1.076(0.324)
	Our Model	0.180(0.080)	**0.703(0.108)**	**0.731(0.159)**	0.951(0.288)

Table A17: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on the landmark risk set ($t \geq 3$) for the coding domain (149 conversations still at risk at turn 3). Each cell is the mean over 20 random 89/30/30 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.078(0.062)	0.917(0.067)	0.988(0.028)	1.003(0.243)
		Cox PH	0.078(0.067)	0.892(0.095)	0.931(0.111)	1.094(0.306)
		Ours	0.073(0.067)	**0.878(0.112)**	**0.888(0.118)**	1.004(0.237)
	without H	Weibull AFT	0.038(0.054)	0.962(0.054)	1.000(0.000)	1.205(0.283)
		Cox PH	0.053(0.078)	0.940(0.090)	0.985(0.047)	1.128(0.310)
		Ours	0.067(0.082)	**0.920(0.102)**	**0.969(0.080)**	1.155(0.298)
0.20	with H	Weibull AFT	0.190(0.087)	0.798(0.091)	0.973(0.057)	0.823(0.301)
		Cox PH	0.163(0.094)	0.745(0.140)	0.788(0.181)	0.931(0.235)
		Ours	0.180(0.080)	**0.703(0.108)**	**0.731(0.159)**	0.951(0.288)
	without H	Weibull AFT	0.178(0.102)	0.820(0.104)	0.996(0.017)	0.931(0.362)
		Cox PH	0.162(0.077)	0.813(0.099)	0.942(0.103)	1.092(0.345)
		Ours	0.162(0.089)	**0.788(0.108)**	**0.885(0.140)**	1.202(0.283)

Table A18: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H on the landmark risk set ($t \geq 3$) for the research domain. Qwen results use 219 conversations and 20 random 131/44/44 conversation-level splits; Llama results use 114 conversations and 20 random 68/23/23 conversation-level splits. Each cell reports the across-split mean, with the standard deviation in parentheses.

LLM	Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
Qwen	0.10	with H	Weibull AFT	0.077(0.051)	0.918(0.060)	0.994(0.029)	0.985(0.257)
			Cox PH	0.083(0.051)	0.826(0.145)	0.871(0.165)	1.108(0.295)
			Ours	0.084(0.049)	**0.787(0.143)**	**0.818(0.173)**	1.135(0.328)
		without H	Weibull AFT	0.065(0.043)	0.933(0.046)	0.997(0.014)	0.971(0.293)
			Cox PH	0.053(0.055)	0.912(0.122)	0.952(0.113)	1.120(0.327)
			Ours	0.056(0.058)	**0.898(0.135)**	**0.934(0.127)**	1.162(0.337)
	0.20	with H	Weibull AFT	0.195(0.060)	0.788(0.103)	0.976(0.108)	0.447(0.336)
			Cox PH	0.174(0.059)	0.517(0.182)	0.561(0.199)	1.175(0.679)
			Ours	0.180(0.059)	**0.477(0.180)**	**0.513(0.197)**	1.238(0.681)
		without H	Weibull AFT	0.197(0.057)	0.789(0.092)	0.979(0.094)	0.488(0.329)
			Cox PH	0.155(0.071)	0.544(0.191)	0.573(0.201)	1.195(0.443)
			Ours	0.159(0.069)	**0.537(0.192)**	**0.569(0.203)**	1.157(0.489)
Llama	0.10	with H	Weibull AFT	0.067(0.065)	0.933(0.065)	1.000(0.000)	1.037(0.325)
			Cox PH	0.057(0.058)	0.924(0.091)	0.950(0.155)	1.072(0.289)
			Ours	0.072(0.083)	**0.896(0.130)**	**0.915(0.177)**	1.103(0.304)
		without H	Weibull AFT	0.015(0.047)	**0.985(0.047)**	1.000(0.000)	1.145(0.298)
			Cox PH	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.161(0.293)
			Ours	0.007(0.029)	0.993(0.029)[‡]	1.000(0.000)	1.161(0.292)
	0.20	with H	Weibull AFT	0.172(0.080)	0.815(0.111)	0.967(0.149)	0.943(0.347)
			Cox PH	0.161(0.116)	0.791(0.162)	0.869(0.213)	0.914(0.377)
			Ours	0.157(0.118)	**0.776(0.171)**	**0.815(0.234)**	0.964(0.321)
		without H	Weibull AFT	0.135(0.101)	0.852(0.130)	0.967(0.149)	1.017(0.240)
			Cox PH	0.102(0.135)	0.865(0.194)	0.914(0.212)	1.080(0.417)
			Ours	0.161(0.145)	**0.800(0.179)**	**0.894(0.149)**	1.074(0.265)

Table A19: Conformal warning results on the landmark risk set ($t \geq 3$) for the research domain. Qwen results use 219 conversations still at risk at turn 3 and 20 random 131/44/44 conversation-level splits; Llama results use 114 conversations still at risk at turn 3 and 20 random 68/23/23 conversation-level splits. Each cell reports the across-split mean, with the standard deviation in parentheses.

LLM	Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
Qwen	0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.060(0.056)	0.884(0.098)	0.921(0.070)	1.154(0.305)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.011(0.038)	0.989(0.038)	1.000(0.000)	1.285(0.307)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.056(0.058)	0.898(0.135)	0.934(0.127)	1.162(0.337)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.078(0.052)	0.817(0.177)	0.852(0.198)	1.108(0.384)
		$R_{i,t} = q_{i,t}$	0.053(0.055)	0.898(0.097)	0.931(0.068)	1.165(0.319)
		$R_{i,t} = H_{i,t}$	0.073(0.037)	0.831(0.090)	0.863(0.095)	1.300(0.301)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.083(0.059)	0.814(0.112)	0.853(0.093)	1.218(0.326)
		Our Model	0.084(0.049)	**0.787(0.143)**	**0.818(0.173)**	1.135(0.328)
	0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.125(0.047)	0.694(0.078)	0.744(0.077)	1.185(0.342)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.058(0.057)	0.942(0.057)	1.000(0.000)	1.054(0.395)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.159(0.069)	0.537(0.192)	0.569(0.203)	1.157(0.489)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.178(0.060)	0.483(0.172)	0.519(0.183)	1.156(0.721)
		$R_{i,t} = q_{i,t}$	0.125(0.047)	0.699(0.087)	0.750(0.088)	1.191(0.345)
		$R_{i,t} = H_{i,t}$	0.177(0.060)	0.483(0.177)	0.518(0.184)	1.425(0.477)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.178(0.064)	0.497(0.181)	0.539(0.181)	1.427(0.364)
		Our Model	0.180(0.059)	**0.477(0.180)**	**0.513(0.197)**	1.238(0.681)
Llama	0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.161(0.293)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.161(0.293)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.007(0.029)	0.993(0.029)[‡]	1.000(0.000)	1.161(0.292)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.035(0.071)	0.950(0.136)	0.961(0.174)	1.055(0.323)
		$R_{i,t} = q_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.161(0.293)
		$R_{i,t} = H_{i,t}$	0.022(0.054)	0.965(0.096)	0.967(0.149)	1.110(0.306)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.050(0.078)	0.920(0.124)	0.921(0.171)	1.106(0.280)
		Our Model	0.072(0.083)	**0.896(0.130)**	**0.915(0.177)**	1.103(0.304)
	0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.052(0.111)	0.913(0.183)	0.908(0.202)	1.141(0.298)
		$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.143(0.284)
		$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.161(0.145)	0.800(0.179)	0.894(0.149)	1.074(0.265)
		$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.111(0.103)	0.843(0.148)	0.875(0.174)	0.956(0.289)
		$R_{i,t} = q_{i,t}$	0.052(0.111)	0.913(0.183)	0.908(0.202)	1.141(0.298)
		$R_{i,t} = H_{i,t}$	0.111(0.112)	0.835(0.145)	0.851(0.142)	0.960(0.290)
		$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.204(0.136)	0.672(0.178)	0.662(0.198)	0.907(0.361)
		Our Model	0.157(0.118)	**0.776(0.171)**	**0.815(0.234)**	0.964(0.321)

C.4 RESULTS ON THE FULL CONVERSATIONS

Table A20: Turn-level prediction on the full risk set for the medical domain. For Qwen, results are based on 250 conversations (1017 at-risk turns) with 150/50/50 train/validation/test conversation-level splits. For Llama, results are based on 249 conversations with 149/50/50 splits. Values are means over 20 random splits, with across-split standard deviations in parentheses. The mean test-fold event prevalences are 0.1455 for Qwen and 0.3165 for Llama.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\text{nli,mm},i,t}$	0.6290(0.0517)	0.2093(0.0379)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5797(0.0550)	0.1768(0.0292)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t}\}$	0.6359(0.0547)	0.2310(0.0462)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\text{nli},i,t}\}$	0.5888(0.0518)	0.1844(0.0293)
	$R_{i,t} = q_{i,t}$	0.6220(0.0500)	0.2018(0.0343)
	$R_{i,t} = H_{i,t}$	0.4885(0.0433)	0.1654(0.0211)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.6175(0.0388)	0.2040(0.0321)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t} + H_{\text{nli},i,t}\}$	**0.6360(0.0541)**	**0.2355(0.0462)**
Llama	$R_{i,t} = U_{\text{nli,mm},i,t}$	0.5240(0.0456)	0.3241(0.0238)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.6208(0.0460)	0.3971(0.0546)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t}\}$	0.6189(0.0445)	0.4058(0.0503)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\text{nli},i,t}\}$	0.6171(0.0503)	0.3994(0.0561)
	$R_{i,t} = q_{i,t}$	0.5251(0.0450)	0.3285(0.0217)
	$R_{i,t} = H_{i,t}$	0.4152(0.0481)	0.3040(0.0222)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4845(0.0607)	0.3116(0.0265)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\text{nli,mm},i,t} + H_{\text{nli},i,t}\}$	**0.6227(0.0450)**	**0.4143(0.0486)**

Table A21: Turn-level prediction across hazard-model benchmarks on the full risk set for the medical domain. For Qwen, results are based on 250 conversations (1017 at-risk turns) with 150/50/50 train/validation/test conversation-level splits. For Llama, results are based on 249 conversations with 149/50/50 splits. Values are means over 20 random splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.5972(0.0457)	0.1987(0.0251)
		Cox PH	0.6291(0.0524)	0.2260(0.0428)
		Ours	**0.6359(0.0547)**	**0.2310(0.0462)**
	with H	Weibull AFT	0.6021(0.0429)	0.1995(0.0284)
		Cox PH	0.6295(0.0552)	0.2279(0.0430)
		Ours	**0.6360(0.0541)**	**0.2355(0.0462)**
Llama	without H	Weibull AFT	0.5673(0.0536)	0.3612(0.0518)
		Cox PH	0.6078(0.0504)	0.3828(0.0615)
		Ours	**0.6189(0.0445)**	**0.4058(0.0503)**
	with H	Weibull AFT	0.5579(0.0566)	0.3563(0.0562)
		Cox PH	0.6058(0.0541)	0.3834(0.0608)
		Ours	**0.6227(0.0450)**	**0.4143(0.0486)**

Table A22: Turn-level prediction on the full risk set. For Qwen, results are based on 250 conversations (863 at-risk turns) with 150/50/50 train/validation/test conversation-level splits. For Llama, results are based on 249 conversations with 149/50/50 splits. Values are means over 20 random splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.5792(0.0522)	0.2639(0.0402)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5787(0.0508)	0.2600(0.0400)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.5999(0.0528)	0.2927(0.0481)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.6005(0.0509)	0.2766(0.0378)
	$R_{i,t} = q_{i,t}$	0.5788(0.0490)	0.2609(0.0354)
	$R_{i,t} = H_{i,t}$	0.4815(0.0441)	0.2460(0.0244)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5631(0.0441)	0.2818(0.0402)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6046(0.0546)**	**0.2973(0.0510)**
Llama	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.5860(0.0605)	0.6367(0.0803)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.6265(0.0508)	0.6557(0.0659)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.6291(0.0489)	0.6739(0.0626)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.6210(0.0481)	0.6558(0.0660)
	$R_{i,t} = q_{i,t}$	0.5874(0.0589)	0.6366(0.0753)
	$R_{i,t} = H_{i,t}$	0.3895(0.0514)	0.5507(0.0674)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5177(0.0586)	0.5961(0.0790)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6371(0.0500)**	**0.6900(0.0639)**

Table A23: Turn-level prediction across hazard-model benchmarks on the full risk set. For Qwen, results are based on 250 conversations with 150/50/50 splits; for Llama, results are based on 249 conversations with 149/50/50 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.5376(0.0505)	0.2500(0.0373)
		Cox PH	0.5995(0.0650)	0.2922(0.0626)
		Ours	**0.5999(0.0528)**	**0.2927(0.0481)**
	with H	Weibull AFT	0.5413(0.0534)	0.2520(0.0362)
		Cox PH	0.5918(0.0656)	0.2883(0.0607)
		Ours	**0.6046(0.0546)**	**0.2973(0.0510)**
Llama	without H	Weibull AFT	0.5901(0.0575)	0.6415(0.0660)
		Cox PH	0.6112(0.0628)	0.6538(0.0705)
		Ours	**0.6291(0.0489)**	**0.6739(0.0626)**
	with H	Weibull AFT	0.5871(0.0543)	0.6401(0.0634)
		Cox PH	0.6117(0.0602)	0.6560(0.0670)
		Ours	**0.6371(0.0500)**	**0.6900(0.0639)**

Table A24: Turn-level prediction on the full risk set for the coding domain. For Qwen, results are based on 299 conversations (933 at-risk turns) with 179/60/60 train/validation/test conversation-level splits. For Llama, results are based on 296 conversations with 178/59/59 splits. Values are means over 20 random splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.5504(0.0449)	0.2803(0.0336)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5719(0.0348)	0.2945(0.0341)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.5709(0.0346)	0.3021(0.0366)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5719(0.0340)	0.2993(0.0352)
	$R_{i,t} = q_{i,t}$	0.5519(0.0452)	0.2803(0.0331)
	$R_{i,t} = H_{i,t}$	0.4342(0.0404)	0.2370(0.0296)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4805(0.0398)	0.2657(0.0370)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.5728(0.0333)**	**0.3052(0.0347)**
Llama	$R_{i,t} = U_{\mathrm{nli},\mathrm{mm},i,t}$	0.4927(0.0379)	0.5845(0.0526)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5103(0.0471)	0.5912(0.0556)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t}\}$	0.5095(0.0490)	0.5937(0.0613)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5101(0.0354)	0.5965(0.0498)
	$R_{i,t} = q_{i,t}$	0.4915(0.0388)	0.5799(0.0519)
	$R_{i,t} = H_{i,t}$	0.4791(0.0512)	0.5756(0.0653)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.4628(0.0533)	0.5835(0.0644)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli},\mathrm{mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.5106(0.0413)**	**0.5972(0.0427)**

Table A25: Turn-level prediction across hazard-model benchmarks on the full risk set for the coding domain. For Qwen, results are based on 299 conversations with 179/60/60 splits; for Llama, results are based on 296 conversations with 178/59/59 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.4811(0.0496)	0.2513(0.0231)
		Cox PH	0.5601(0.0342)	0.2949(0.0361)
		Ours	**0.5709(0.0346)**	**0.3021(0.0366)**
	with H	Weibull AFT	0.5031(0.0483)	0.2740(0.0283)
		Cox PH	0.5657(0.0338)	**0.3096(0.0385)**
		Ours	**0.5728(0.0333)**	0.3052(0.0347)
Llama	without H	Weibull AFT	0.4754(0.0401)	0.5765(0.0571)
		Cox PH	0.4953(0.0600)	0.5917(0.0614)
		Ours	**0.5095(0.0490)**	**0.5937(0.0613)**
	with H	Weibull AFT	0.4933(0.0420)	0.5959(0.0580)
		Cox PH	0.5011(0.0615)	0.5963(0.0643)
		Ours	**0.5106(0.0413)**	**0.5972(0.0427)**

Table A26: Turn-level prediction on the full risk set for the research domain. For Qwen, results are based on 250 conversations (1249 at-risk turns) with 150/50/50 train/validation/test conversation-level splits. For Llama, results are based on 250 conversations (738 at-risk turns) with 150/50/50 splits. Values are means over 20 random splits, with across-split standard deviations in parentheses.

LLM	Score construction	AUROC ↑	AUPR ↑
Qwen	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.5094(0.0501)	0.0798(0.0098)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	**0.5288(0.0666)**	0.0870(0.0142)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.5163(0.0618)	0.0869(0.0149)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5138(0.0598)	0.0979(0.0280)
	$R_{i,t} = q_{i,t}$	0.5077(0.0503)	0.0794(0.0098)
	$R_{i,t} = H_{i,t}$	0.5145(0.0541)	0.0823(0.0135)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.5143(0.0540)	0.0820(0.0165)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	0.5190(0.0658)	**0.1052(0.0323)**
Llama	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.6138(0.0392)	0.3692(0.0430)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.5427(0.0490)	0.3110(0.0350)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.6180(0.0395)	0.3897(0.0519)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.5741(0.0435)	0.3539(0.0492)
	$R_{i,t} = q_{i,t}$	0.6128(0.0400)	0.3687(0.0431)
	$R_{i,t} = H_{i,t}$	0.5240(0.0592)	0.3378(0.0487)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.6198(0.0532)	0.3985(0.0638)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$	**0.6222(0.0431)**	**0.3999(0.0643)**

Table A27: Turn-level prediction across hazard-model benchmarks on the full risk set for the research domain. For Qwen, results are based on 250 conversations with 150/50/50 splits; for Llama, results are based on 250 conversations with 150/50/50 splits. Values are means over 20 random conversation-level splits, with across-split standard deviations in parentheses.

LLM	Historical feature	Method	AUROC ↑	AUPR ↑
Qwen	without H	Weibull AFT	0.5105(0.0468)	0.0823(0.0101)
		Cox PH	0.4654(0.0524)	0.0732(0.0082)
		Ours	**0.5163(0.0618)**	**0.0869(0.0149)**
	with H	Weibull AFT	0.5073(0.0511)	0.0833(0.0124)
		Cox PH	0.4649(0.0554)	0.0773(0.0158)
		Ours	**0.5190(0.0658)**	**0.1052(0.0323)**
Llama	without H	Weibull AFT	0.5787(0.0343)	0.3581(0.0359)
		Cox PH	0.6052(0.0441)	0.3885(0.0534)
		Ours	**0.6180(0.0395)**	**0.3897(0.0519)**
	with H	Weibull AFT	0.5870(0.0434)	0.3639(0.0398)
		Cox PH	0.6102(0.0421)	0.3964(0.0615)
		Ours	**0.6222(0.0431)**	**0.3999(0.0643)**

C.4.1 CONFORMAL RESULTS

Table A28: Conformal warning results on the full risk set ($t \geq 1$) for the medical domain. Results are based on all 250 Qwen conversations. Each cell reports the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.088(0.063)	0.824(0.101)	0.790(0.109)	1.581(0.308)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.745(0.292)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.092(0.064)	0.799(0.105)	0.740(0.122)	1.600(0.321)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.745(0.292)
	$R_{i,t} = q_{i,t}$	0.065(0.060)	0.880(0.097)	0.869(0.099)	1.578(0.306)
	$R_{i,t} = H_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.745(0.292)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.091(0.058)	0.829(0.091)	0.810(0.096)	1.563(0.282)
	Our Model	0.098(0.069)	**0.786(0.106)**	**0.724(0.113)**	1.581(0.317)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.169(0.057)	0.671(0.081)	0.619(0.085)	1.364(0.349)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.010(0.045)	0.990(0.045)[‡]	1.000(0.000)	1.751(0.302)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.192(0.061)	0.636(0.083)	0.590(0.091)	1.407(0.372)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.011(0.049)	0.987(0.058)	0.995(0.021)	1.755(0.309)
	$R_{i,t} = q_{i,t}$	0.138(0.053)	0.734(0.069)	0.695(0.078)	1.522(0.309)
	$R_{i,t} = H_{i,t}$	0.100(0.120)	0.878(0.149)	0.948(0.084)	1.620(0.332)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.174(0.054)	0.674(0.084)	0.638(0.099)	1.263(0.299)
	Our Model	0.192(0.063)	**0.627(0.094)**	**0.569(0.104)**	1.407(0.349)

Table A29: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H on the full risk set ($t \geq 1$) for the medical domain (250 Qwen conversations). Each cell reports the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.089(0.053)	0.851(0.073)	0.857(0.122)	1.591(0.319)
		Cox PH	0.095(0.065)	0.797(0.101)	0.743(0.111)	1.564(0.330)
		Ours	0.098(0.069)	**0.786(0.106)**	**0.724(0.113)**	1.581(0.317)
	without H	Weibull AFT	0.092(0.059)	0.837(0.087)	0.831(0.125)	1.547(0.350)
		Cox PH	0.093(0.065)	0.805(0.096)	0.757(0.091)	1.541(0.337)
		Ours	0.092(0.064)	**0.799(0.105)**	**0.740(0.122)**	1.600(0.321)
0.20	with H	Weibull AFT	0.185(0.070)	0.668(0.117)	0.650(0.179)	1.344(0.397)
		Cox PH	0.185(0.067)	0.645(0.096)	0.595(0.114)	1.482(0.331)
		Ours	0.192(0.063)	**0.627(0.094)**	**0.569(0.104)**	1.407(0.349)
	without H	Weibull AFT	0.179(0.069)	0.679(0.108)	0.662(0.156)	1.344(0.378)
		Cox PH	0.187(0.069)	0.643(0.100)	0.595(0.124)	1.440(0.328)
		Ours	0.192(0.061)	**0.636(0.083)**	**0.590(0.091)**	1.407(0.372)

Table A30: Conformal warning results on the full risk set ($t \geq 1$) for the legal domain. Results are based on all 250 Qwen conversations. Each cell reports the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.068(0.041)	0.930(0.041)	0.993(0.022)	1.490(0.227)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.074(0.067)	0.909(0.079)	**0.939(0.074)**	1.500(0.234)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = q_{i,t}$	0.068(0.041)	0.930(0.041)	0.993(0.022)	1.490(0.227)
	$R_{i,t} = H_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.077(0.063)	0.920(0.066)	0.989(0.026)	1.499(0.224)
	Our Model	0.079(0.071)	**0.904(0.084)**	**0.939(0.074)**	1.508(0.238)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.149(0.082)	0.834(0.093)	0.939(0.074)	1.479(0.251)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.156(0.084)	0.801(0.102)	**0.846(0.104)**	1.500(0.268)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = q_{i,t}$	0.144(0.080)	0.839(0.089)	0.939(0.074)	1.488(0.248)
	$R_{i,t} = H_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.476(0.239)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.174(0.073)	0.808(0.079)	0.936(0.056)	1.471(0.256)
	Our Model	0.168(0.085)	**0.789(0.103)**	**0.846(0.112)**	1.517(0.266)

Our Model denotes $R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t} + H_{\mathrm{nli},i,t}\}$. Bold marks the best $\mathbb{E}[\mathrm{WR}]$ or $\mathbb{E}[\mathrm{UWR}]$ among constructions meeting $\mathbb{E}[\mathrm{MWR}] \leq \alpha$, ties included; underlined $\mathbb{E}[\mathrm{MWR}]$ exceeds α, so that construction buys its warning rate by missing failures and is excluded from the bolding. [‡] warns on essentially every test conversation, so its MWR is degenerate rather than informative and it is likewise excluded. [§] Mean lead is $Y_i - \tau_{\mathrm{warn},i}$, averaged over caught failures only.

Table A31: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on the full risk set ($t \geq 1$) for the legal domain (250 conversations). Each cell is the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Historical feature	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.076(0.065)	0.923(0.065)	0.996(0.016)	1.481(0.207)
		Cox PH	0.062(0.073)	0.926(0.088)	0.957(0.075)	1.484(0.241)
		Ours	0.079(0.071)	**0.904(0.084)**	**0.939(0.074)**	1.508(0.238)
	without H	Weibull AFT	0.077(0.063)	0.922(0.063)	0.996(0.016)	1.490(0.226)
		Cox PH	0.067(0.067)	0.917(0.081)	0.943(0.085)	1.499(0.238)
		Ours	0.074(0.067)	**0.909(0.079)**	**0.939(0.074)**	1.500(0.234)
0.20	with H	Weibull AFT	0.173(0.072)	0.823(0.074)	0.986(0.037)	1.404(0.236)
		Cox PH	0.142(0.074)	0.823(0.092)	0.875(0.108)	1.515(0.252)
		Ours	0.168(0.085)	**0.789(0.103)**	**0.846(0.112)**	1.517(0.266)
	without H	Weibull AFT	0.170(0.076)	0.825(0.075)	0.982(0.046)	1.429(0.257)
		Cox PH	0.142(0.060)	0.822(0.076)	0.871(0.105)	1.512(0.257)
		Ours	0.156(0.084)	**0.801(0.102)**	**0.846(0.104)**	1.500(0.268)

Table A32: Conformal warning results on the full risk set ($t \geq 1$) for the coding domain. Results are based on all 299 Qwen conversations. Each cell reports the mean over 20 random 179/60/60 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.087(0.044)	0.913(0.044)	1.000(0.000)	1.338(0.183)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.081(0.064)	0.914(0.069)	0.977(0.044)	1.351(0.175)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = q_{i,t}$	0.067(0.031)	0.933(0.031)	1.000(0.000)	1.340(0.176)
	$R_{i,t} = H_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.079(0.044)	0.921(0.044)	1.000(0.000)	1.329(0.162)
	Our Model	0.082(0.052)	**0.911(0.056)**	**0.969(0.046)**	1.360(0.174)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.149(0.092)	0.846(0.098)	0.977(0.044)	1.361(0.175)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.115(0.089)	0.873(0.101)	0.942(0.078)	1.366(0.186)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = q_{i,t}$	0.117(0.105)	0.879(0.111)	0.985(0.040)	1.352(0.166)
	$R_{i,t} = H_{i,t}$	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.130(0.063)	0.868(0.064)	0.992(0.024)	1.397(0.258)
	Our Model	0.139(0.107)	**0.843(0.123)**	**0.919(0.095)**	1.377(0.177)

Table A33: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on the full risk set ($t \geq 1$) for the coding domain (299 conversations). Each cell is the mean over 20 random 179/60/60 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
		Cox PH	0.073(0.046)	0.922(0.049)	0.977(0.044)	1.336(0.187)
		Ours	0.082(0.052)	**0.911(0.056)**	**0.969(0.046)**	1.360(0.174)
	without H	Weibull AFT	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
		Cox PH	0.077(0.064)	0.918(0.068)	0.981(0.042)	1.343(0.179)
		Ours	0.081(0.064)	**0.914(0.069)**	**0.977(0.044)**	1.351(0.175)
0.20	with H	Weibull AFT	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
		Cox PH	0.119(0.096)	0.869(0.109)	0.946(0.083)	1.349(0.176)
		Ours	0.139(0.107)	**0.843(0.123)**	**0.919(0.095)**	1.377(0.177)
	without H	Weibull AFT	0.000(0.000)	1.000(0.000)[‡]	1.000(0.000)	1.303(0.163)
		Cox PH	0.105(0.092)	0.883(0.104)	0.946(0.079)	1.356(0.175)
		Ours	0.115(0.089)	**0.873(0.101)**	**0.942(0.078)**	1.366(0.186)

Table A34: Conformal warning results on the full risk set ($t \geq 1$) for the research domain. Results are based on all 250 Qwen conversations. Each cell reports the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.046(0.026)	0.888(0.039)	0.894(0.052)	2.148(0.453)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.065(0.046)	0.935(0.046)	1.000(0.000)	1.812(0.373)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.069(0.048)	0.925(0.053)	0.990(0.018)	1.791(0.373)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.075(0.063)	0.919(0.067)	0.990(0.030)	1.939(0.531)
	$R_{i,t} = q_{i,t}$	0.046(0.026)	0.888(0.039)	0.894(0.052)	2.151(0.451)
	$R_{i,t} = H_{i,t}$	0.077(0.041)	0.881(0.066)	0.932(0.056)	1.807(0.384)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.060(0.048)	0.891(0.066)	0.921(0.048)	1.898(0.419)
	Our Model	0.089(0.067)	**0.837(0.159)**	**0.881(0.156)**	1.880(0.303)
0.20	$R_{i,t} = U_{\mathrm{nli,mm},i,t}$	0.047(0.025)	0.882(0.029)	0.885(0.047)	2.145(0.455)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.139(0.064)	0.861(0.064)	1.000(0.000)	1.223(0.361)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{nli,mm},i,t}\}$	0.158(0.061)	0.768(0.109)	0.881(0.116)	1.313(0.419)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{nli},i,t}\}$	0.178(0.060)	0.724(0.128)	0.842(0.143)	1.235(0.494)
	$R_{i,t} = q_{i,t}$	0.047(0.025)	0.882(0.029)	0.885(0.047)	2.145(0.455)
	$R_{i,t} = H_{i,t}$	0.185(0.065)	0.718(0.121)	0.844(0.121)	1.324(0.548)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.154(0.047)	0.697(0.107)	0.760(0.107)	1.686(0.492)
	Our Model	0.154(0.047)	**0.684(0.064)**	**0.739(0.063)**	1.731(0.428)

Table A35: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H on the full risk set ($t \geq 1$) for the research domain (250 Qwen conversations). Each cell reports the mean over 20 random 150/50/50 conversation-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.082(0.054)	0.918(0.054)	1.000(0.000)	1.692(0.455)
		Cox PH	0.066(0.056)	0.929(0.057)	0.992(0.029)	1.986(0.438)
		Ours	0.089(0.067)	**0.837(0.159)**	**0.881(0.156)**	1.880(0.303)
	without H	Weibull AFT	0.077(0.050)	**0.923(0.050)**	1.000(0.000)	1.725(0.433)
		Cox PH	0.061(0.047)	0.938(0.046)	0.998(0.007)	1.950(0.385)
		Ours	0.069(0.048)	0.925(0.053)	**0.990(0.018)**	1.791(0.373)
0.20	with H	Weibull AFT	0.181(0.068)	0.816(0.068)	0.995(0.012)	1.166(0.503)
		Cox PH	0.198(0.064)	0.722(0.131)	0.871(0.144)	1.346(0.483)
		Ours	0.154(0.047)	**0.684(0.064)**	**0.739(0.063)**	1.731(0.428)
	without H	Weibull AFT	0.182(0.064)	0.815(0.064)	0.995(0.022)	1.150(0.366)
		Cox PH	0.167(0.078)	0.783(0.142)	0.919(0.144)	1.353(0.539)
		Ours	0.158(0.061)	**0.768(0.109)**	**0.881(0.116)**	1.313(0.419)

C.5 Results on PRM800K

Event prevalence in the held-out rows is $\pi = 0.0347$ (≈67200 rows, 2333 events) for a minimum reference pool of 3 and $\pi = 0.0370$ (≈73100 rows, 2718 events) for a pool of 2; these are the AUPR of a random ranker.

Table A36: LOO step-level prediction on PRM800K with $5 \leq K_p \leq 20$. AUROC and AUPR are computed on at-risk rows using the first-failure label $Z_{i,t} = \mathbf{1}\{Y_i = t, \Delta_i = 1\}$. Mean over 20 random problem-level 60/20/20 splits, across-split standard deviation in parentheses.

Score construction	Min. reference pool = 3		Min. reference pool = 2	
	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
$R_{i,t} = U_{\mathrm{sem},i,t}$	0.4408(0.0059)	0.0294(0.0007)	0.4225(0.0055)	0.0305(0.0007)
$R_{i,t} = q_{\mathrm{sem},i,t}$	0.4650(0.0063)	0.0311(0.0009)	0.4431(0.0059)	0.0321(0.0008)
$R_{i,t} = H_{\mathrm{sem},i,t}$	0.5977(0.0053)	0.0418(0.0013)	0.5772(0.0053)	0.0425(0.0012)
$R_{i,t} = \frac{1}{2}(\widetilde{U}_{\mathrm{sem},i,t} + \widetilde{H}_{\mathrm{sem},i,t})$	0.5666(0.0043)	0.0377(0.0010)	0.5430(0.0048)	0.0384(0.0009)
$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.6571(0.0038)	0.0518(0.0012)	0.6527(0.0036)	0.0543(0.0012)
$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.6568(0.0039)	0.0530(0.0014)	0.6552(0.0038)	0.0566(0.0013)
$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.6623(0.0046)	0.0537(0.0018)	0.6555(0.0041)	0.0555(0.0015)
Our Model	**0.6696(0.0045)**	**0.0578(0.0018)**	**0.6653(0.0043)**	**0.0607(0.0015)**

Table A37: Turn-level prediction across hazard-model benchmarks on PRM800K with $5 \leq K_p \leq 20$. Mean over the same 20 random problem-level 60/20/20 splits, across-split standard deviation in parentheses.

Historical feature	Method	Min. reference pool = 3		Min. reference pool = 2	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
without H	Weibull AFT	0.6492(0.0063)	0.0502(0.0021)	0.6459(0.0056)	0.0529(0.0020)
	Cox PH	0.6381(0.0083)	0.0476(0.0021)	0.6456(0.0081)	0.0545(0.0030)
	Ours	**0.6568(0.0039)**	**0.0530(0.0014)**	**0.6552(0.0038)**	**0.0566(0.0013)**
with H	Weibull AFT	0.6572(0.0072)	0.0545(0.0027)	0.6552(0.0069)	0.0583(0.0024)
	Cox PH	0.6426(0.0093)	0.0487(0.0024)	0.6523(0.0103)	0.0572(0.0038)
	Ours	**0.6696(0.0045)**	**0.0578(0.0018)**	**0.6653(0.0043)**	**0.0607(0.0015)**

Table A38: LOO step-level prediction on PRM800K with $K_p \geq 5$, the uncapped robustness setting. Mean over the same 20 random problem-level 60/20/20 splits, across-split standard deviation in parentheses.

Score construction	Min. reference pool = 3		Min. reference pool = 2	
	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
$R_{i,t} = U_{\mathrm{sem},i,t}$	0.5345(0.0383)	0.0353(0.0023)	0.5226(0.0370)	0.0360(0.0022)
$R_{i,t} = q_{\mathrm{sem},i,t}$	0.5332(0.0375)	0.0334(0.0010)	0.5211(0.0366)	0.0342(0.0010)
$R_{i,t} = H_{\mathrm{sem},i,t}$	0.6198(0.0260)	0.0426(0.0012)	0.6095(0.0259)	0.0434(0.0012)
$R_{i,t} = \frac{1}{2}(\widetilde{U}_{\mathrm{sem},i,t} + \widetilde{H}_{\mathrm{sem},i,t})$	0.5977(0.0315)	0.0391(0.0010)	0.5883(0.0306)	0.0400(0.0010)
$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.6454(0.0123)	0.0478(0.0067)	0.6435(0.0110)	0.0494(0.0059)
$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.6699(0.0044)	0.0498(0.0022)	0.6609(0.0046)	0.0511(0.0021)
$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.6647(0.0075)	0.0506(0.0024)	0.6625(0.0064)	0.0527(0.0021)
Our Model	**0.6774(0.0161)**	**0.0531(0.0026)**	**0.6685(0.0157)**	**0.0535(0.0023)**

Table A39: Turn-level prediction across hazard-model benchmarks on PRM800K with $K_p \geq 5$. Mean over 20 problem-level 60/20/20 splits, across-split standard deviation in parentheses.

Historical feature	Method	Min. reference pool = 3		Min. reference pool = 2	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
without H	Weibull AFT	0.6454(0.0135)	0.0480(0.0025)	0.6110(0.0621)	0.0459(0.0079)
	Cox PH	0.5804(0.0253)	0.0402(0.0028)	0.5684(0.0260)	0.0411(0.0026)
	Ours	**0.6699(0.0044)**	**0.0498(0.0022)**	**0.6609(0.0046)**	**0.0511(0.0021)**
with H	Weibull AFT	0.6456(0.0115)	0.0474(0.0026)	0.6407(0.0114)	0.0488(0.0022)
	Cox PH	0.5982(0.0229)	0.0418(0.0024)	0.5886(0.0234)	0.0425(0.0021)
	Ours	**0.6774(0.0161)**	**0.0531(0.0026)**	**0.6685(0.0157)**	**0.0535(0.0023)**

C.5.1 Conformal Results

Table A40: Conformal warning results across score constructions on PRM800K leave-one-chain-out, $5 \leq K_p \leq 20$, minimum reference pool 2. Each cell reports the mean over 20 random 60/20/20 problem-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0977(0.009)	0.5852(0.030)	0.5661(0.030)	6.604(0.122)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0864(0.003)	0.5610(0.006)	0.5176(0.006)	4.524(0.087)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.1004(0.005)	0.5169(0.017)	0.4764(0.016)	4.357(0.094)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.1017(0.004)	0.5020(0.015)	0.4577(0.016)	4.620(0.081)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0928(0.008)	0.5996(0.030)	0.5791(0.030)	6.403(0.098)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.1006(0.007)	0.5309(0.024)	0.4958(0.024)	5.692(0.115)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.1008(0.005)	0.5290(0.017)	0.4935(0.018)	5.833(0.120)
	Our Model	0.0996(0.005)	**0.5115(0.014)**	**0.4679(0.014)**	4.284(0.104)
0.20	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.1941(0.005)	0.2456(0.019)	0.2333(0.020)	6.434(0.203)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.1906(0.011)	0.2327(0.036)	0.2110(0.034)	4.340(0.221)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.1988(0.006)	0.2102(0.017)	0.1914(0.016)	4.062(0.257)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.2036(0.006)	0.1886(0.015)	0.1656(0.013)	4.631(0.246)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.1892(0.008)	0.2598(0.027)	0.2461(0.026)	6.252(0.191)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.1988(0.008)	0.2194(0.021)	0.2040(0.019)	5.614(0.200)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.2003(0.008)	0.2132(0.023)	0.1975(0.021)	5.923(0.210)
	Our Model	0.1994(0.005)	**0.1995(0.010)**	**0.1776(0.009)**	3.993(0.212)

Table A41: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on PRM800K leave-one-chain-out, $5 \leq K_p \leq 20$, minimum reference pool 2. Each cell is the mean over 20 random 60/20/20 problem-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.1017(0.0047)	0.5051(0.0177)	0.4620(0.0185)	4.695(0.103)
		Cox PH	0.1003(0.0046)	0.5088(0.0168)	0.4652(0.0160)	4.291(0.104)
		Ours	0.0996(0.0048)	**0.5115(0.0145)**	**0.4679(0.0137)**	4.284(0.104)
	without H	Weibull AFT	0.1006(0.0059)	0.5123(0.0196)	0.4703(0.0190)	4.636(0.085)
		Cox PH	0.1004(0.0051)	0.5176(0.0174)	0.4774(0.0168)	4.346(0.095)
		Ours	0.1004(0.0051)	0.5169(0.0169)	0.4764(0.0162)	4.357(0.094)
0.20	with H	Weibull AFT	0.2029(0.0059)	0.1892(0.0144)	0.1674(0.0131)	4.831(0.288)
		Cox PH	0.1994(0.0048)	0.2008(0.0109)	0.1793(0.0101)	3.843(0.160)
		Ours	0.1994(0.0050)	**0.1995(0.0103)**	**0.1776(0.0094)**	3.993(0.212)
	without H	Weibull AFT	0.2005(0.0055)	0.1986(0.0123)	0.1778(0.0116)	4.732(0.270)
		Cox PH	0.1968(0.0038)	0.2266(0.0084)	0.2111(0.0091)	3.888(0.213)
		Ours	0.1988(0.0059)	**0.2102(0.0166)**	**0.1914(0.0159)**	4.062(0.257)

Table A42: Conformal warning results across score constructions on PRM800K leave-one-chain-out, $5 \leq K_p \leq 20$, minimum LOO reference pool 3. The reference-pool threshold removes steps whose Each cell reports the mean over 20 random 60/20/20 problem-level splits, with the across-split standard deviation in parentheses.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0949(0.010)	0.5918(0.033)	0.5712(0.032)	6.601(0.137)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0821(0.020)	0.5767(0.071)	0.5332(0.069)	4.659(0.135)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.0958(0.006)	0.5285(0.022)	0.4859(0.022)	4.542(0.087)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.0973(0.004)	0.5155(0.012)	0.4701(0.012)	4.760(0.097)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0949(0.011)	0.5884(0.038)	0.5665(0.038)	6.399(0.148)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.1000(0.008)	0.5321(0.026)	0.4966(0.025)	5.688(0.121)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.1003(0.006)	0.5296(0.020)	0.4934(0.020)	5.820(0.134)
	Our Model	0.0988(0.004)	**0.5108(0.014)**	**0.4658(0.014)**	4.483(0.065)
0.20	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.1880(0.006)	0.2609(0.020)	0.2467(0.018)	6.449(0.176)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.1893(0.004)	0.2285(0.004)	0.2041(0.005)	4.470(0.194)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.1962(0.006)	0.2091(0.013)	0.1870(0.012)	4.405(0.251)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.1971(0.004)	0.1961(0.009)	0.1706(0.009)	4.626(0.232)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.1843(0.005)	0.2706(0.015)	0.2548(0.014)	6.213(0.176)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.1960(0.007)	0.2243(0.018)	0.2077(0.016)	5.558(0.184)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.1959(0.006)	0.2237(0.015)	0.2066(0.012)	5.896(0.223)
	Our Model	0.1989(0.003)	**0.1931(0.008)**	**0.1688(0.007)**	4.186(0.198)

Table A43: Conformal warning comparison across hazard-model benchmarks with and without the historical feature H, on PRM800K leave-one-chain-out, $5 \leq K_p \leq 20$, minimum reference pool 3. Each cell is the mean over 20 random 60/20/20 problem-level splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.10	with H	Weibull AFT	0.0964(0.0034)	0.5210(0.0093)	0.4763(0.0093)	4.901(0.157)
		Cox PH	0.0990(0.0036)	0.5112(0.0125)	0.4666(0.0129)	4.513(0.076)
		Ours	0.0988(0.0042)	**0.5108(0.0141)**	**0.4658(0.0137)**	4.483(0.065)
	without H	Weibull AFT	0.0935(0.0065)	0.5368(0.0198)	0.4940(0.0193)	4.835(0.127)
		Cox PH	0.0944(0.0138)	0.5333(0.0432)	0.4901(0.0409)	4.597(0.144)
		Ours	0.0958(0.0061)	**0.5285(0.0220)**	**0.4859(0.0220)**	4.542(0.087)
0.20	with H	Weibull AFT	0.1972(0.0043)	0.1985(0.0111)	0.1739(0.0104)	5.033(0.340)
		Cox PH	0.1992(0.0062)	0.1948(0.0124)	0.1700(0.0096)	4.081(0.127)
		Ours	0.1989(0.0030)	**0.1931(0.0080)**	**0.1688(0.0068)**	4.186(0.198)
	without H	Weibull AFT	0.1933(0.0050)	0.2132(0.0107)	0.1887(0.0089)	4.901(0.290)
		Cox PH	0.1924(0.0046)	0.2308(0.0124)	0.2116(0.0128)	4.193(0.182)
		Ours	0.1962(0.0061)	**0.2091(0.0128)**	**0.1870(0.0115)**	4.405(0.251)

The unrestricted regime violates the balanced-cluster setting needed for clean chain-level conformal interpretation.

Table A44: Conformal warning results across score constructions on PRM800K leave-one-chain-out, $K_p \geq 5$, with minimum LOO reference pool 2. Each cell is the mean over 20 random splits, with the across-split standard deviation in parentheses

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.05	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0420(0.014)	0.7439(0.060)	0.7294(0.063)	5.004(0.107)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0420(0.008)	0.7227(0.035)	0.7024(0.036)	3.643(0.293)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.0477(0.006)	0.6779(0.022)	0.6570(0.024)	3.661(0.220)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.0484(0.013)	0.6664(0.040)	0.6392(0.041)	3.765(0.212)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0447(0.018)	0.7472(0.078)	0.7369(0.082)	5.060(0.084)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.0473(0.020)	0.6752(0.066)	0.6489(0.064)	4.403(0.158)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.0486(0.019)	0.6716(0.079)	0.6461(0.082)	4.392(0.103)
	Our Model	0.0492(0.013)	**0.6609(0.046)**	**0.6334(0.045)**	3.735(0.207)
0.10	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0813(0.009)	0.4830(0.046)	0.4552(0.045)	5.193(0.081)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0839(0.012)	0.5325(0.052)	0.5203(0.050)	3.119(0.168)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.0908(0.006)	0.4722(0.035)	0.4535(0.034)	3.305(0.081)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.0803(0.007)	0.4734(0.036)	0.4421(0.036)	3.306(0.077)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0736(0.007)	0.5155(0.044)	0.4864(0.045)	5.149(0.123)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.0742(0.007)	0.4783(0.045)	0.4407(0.045)	4.406(0.098)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.0739(0.007)	0.4710(0.042)	0.4311(0.042)	4.545(0.096)
	Our Model	0.0875(0.007)	**0.4365(0.035)**	**0.4049(0.033)**	3.349(0.087)
0.20	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.1928(0.013)	0.0758(0.039)	0.0758(0.038)	3.150(1.465)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.1791(0.013)	0.1610(0.027)	0.1659(0.027)	2.728(0.301)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.1954(0.015)	0.0588(0.027)	0.0575(0.027)	2.591(0.668)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.1958(0.015)	0.0462(0.026)	0.0421(0.024)	3.081(0.411)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.1916(0.015)	0.0600(0.034)	0.0543(0.032)	5.162(0.317)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.1938(0.015)	0.0520(0.031)	0.0470(0.029)	4.405(0.461)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.1944(0.016)	0.0477(0.029)	0.0423(0.027)	4.940(0.486)
	Our Model	0.1960(0.015)	**0.0432(0.021)**	**0.0386(0.020)**	3.557(0.390)

Table A45: Conformal warning comparison across hazard-model benchmarks with on PRM800K leave-one-chain-out, $K_p \geq 5$, with minimum LOO reference pool 2. Each cell is the mean over 20 random splits, with the across-split standard deviation in parentheses.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.05	with H	Weibull AFT	0.0478(0.0136)	0.6760(0.0581)	0.6508(0.0602)	4.017(0.127)
		Cox PH	0.0489(0.0118)	0.6805(0.0492)	0.6576(0.0523)	3.636(0.236)
		Ours	0.0492(0.0131)	**0.6609(0.0460)**	**0.6334(0.0452)**	3.735(0.207)
	without H	Weibull AFT	0.0477(0.0120)	0.6959(0.0546)	0.6758(0.0558)	4.170(0.261)
		Cox PH	0.0483(0.0067)	0.6886(0.0264)	0.6673(0.0279)	3.628(0.232)
		Ours	0.0477(0.0063)	**0.6779(0.0222)**	**0.6570(0.0236)**	3.661(0.220)
0.10	with H	Weibull AFT	0.0857(0.0075)	0.4403(0.0344)	0.4074(0.0341)	4.083(0.201)
		Cox PH	0.0857(0.0055)	0.4579(0.0233)	0.4293(0.0216)	3.383(0.067)
		Ours	0.0875(0.0071)	**0.4365(0.0346)**	**0.4049(0.0335)**	3.349(0.087)
	without H	Weibull AFT	0.0883(0.0095)	**0.4612(0.0500)**	**0.4366(0.0506)**	3.952(0.216)
		Cox PH	0.0868(0.0062)	0.4648(0.0315)	0.4393(0.0318)	3.404(0.059)
		Ours	0.0908(0.0064)	0.4722(0.0346)	0.4535(0.0338)	3.305(0.081)
0.20	with H	Weibull AFT	0.1958(0.0148)	0.0518(0.0267)	0.0492(0.0257)	3.518(0.492)
		Cox PH	0.1964(0.0156)	0.0581(0.0220)	0.0576(0.0238)	3.250(0.677)
		Ours	0.1960(0.0155)	**0.0432(0.0212)**	**0.0386(0.0197)**	3.557(0.390)
	without H	Weibull AFT	0.1943(0.0130)	0.0688(0.0314)	0.0688(0.0328)	2.329(0.841)
		Cox PH	0.1947(0.0135)	0.0733(0.0277)	0.0748(0.0278)	2.633(0.668)
		Ours	0.1954(0.0147)	**0.0588(0.0273)**	**0.0575(0.0273)**	2.591(0.668)

Table A46: Conformal warning results across score constructions on PRM800K leave-one-chain-out, $K_p \geq 5$, with minimum LOO reference pool 3.

Target α	Score construction	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.05	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0434(0.009)	0.6694(0.068)	0.6486(0.073)	4.718(0.136)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0438(0.003)	0.6852(0.018)	0.6683(0.020)	3.256(0.237)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.0483(0.008)	0.6390(0.034)	0.6172(0.034)	3.349(0.128)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.0470(0.010)	0.6413(0.047)	0.6185(0.047)	3.304(0.183)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0410(0.011)	0.7255(0.080)	0.7141(0.085)	4.658(0.112)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.0455(0.013)	0.6380(0.075)	0.6126(0.077)	3.958(0.087)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.0445(0.013)	0.6483(0.099)	0.6238(0.106)	4.040(0.105)
	Our Model	0.0483(0.008)	**0.6326(0.038)**	**0.6094(0.039)**	3.363(0.165)
0.10	$R_{i,t} = U_{\mathrm{sem},i,t}$	0.0800(0.005)	0.4386(0.035)	0.4133(0.035)	4.638(0.098)
	$R_{i,t} = \widehat{\lambda}\{f(t)\}$	0.0857(0.011)	0.4850(0.063)	0.4769(0.063)	2.722(0.278)
	$R_{i,t} = \widehat{\lambda}\{f(t) + U_{\mathrm{sem},i,t}\}$	0.0942(0.004)	0.4003(0.053)	0.3838(0.059)	2.861(0.076)
	$R_{i,t} = \widehat{\lambda}\{f(t) + H_{\mathrm{sem},i,t}\}$	0.0880(0.008)	0.4128(0.048)	0.3914(0.048)	2.829(0.076)
	$R_{i,t} = q_{\mathrm{sem},i,t}$	0.0802(0.007)	0.4366(0.050)	0.4111(0.051)	4.607(0.053)
	$R_{i,t} = H_{\mathrm{sem},i,t}$	0.0839(0.008)	0.3930(0.043)	0.3623(0.042)	3.937(0.085)
	$R_{i,t} = (\widetilde{U}_{i,t} + \widetilde{H}_{i,t})/2$	0.0842(0.007)	0.3824(0.040)	0.3497(0.039)	4.156(0.099)
	Our Model	0.0937(0.005)	**0.3664(0.034)**	**0.3418(0.035)**	2.964(0.107)

Table A47: Conformal warning comparison across hazard-model benchmarks on PRM800K leave-one-chain-out, $K_p \geq 5$, with minimum LOO reference pool 3.

Target α	Setting	Method	$\mathbb{E}[\mathrm{MWR}](\leq \alpha)$	$\mathbb{E}[\mathrm{WR}] \downarrow$	$\mathbb{E}[\mathrm{UWR}] \downarrow$	Mean lead[§]
0.05	with H	Weibull AFT	0.0468(0.0110)	0.6483(0.0768)	0.6268(0.0820)	3.766(0.164)
		Cox PH	0.0472(0.0109)	0.6483(0.0452)	0.6272(0.0446)	3.253(0.141)
		Ours	0.0483(0.0083)	**0.6326(0.0382)**	**0.6094(0.0391)**	3.363(0.165)
	without H	Weibull AFT	0.0467(0.0117)	0.6553(0.0678)	0.6352(0.0702)	3.894(0.204)
		Cox PH	0.0479(0.0096)	0.6439(0.0358)	0.6228(0.0345)	3.223(0.135)
		Ours	0.0483(0.0084)	**0.6390(0.0342)**	**0.6172(0.0337)**	3.349(0.128)
0.10	with H	Weibull AFT	0.0906(0.0049)	0.3902(0.0328)	0.3671(0.0364)	3.438(0.293)
		Cox PH	0.0914(0.0052)	0.3900(0.0359)	0.3679(0.0370)	3.073(0.112)
		Ours	0.0937(0.0053)	**0.3664(0.0339)**	**0.3418(0.0347)**	2.964(0.107)
	without H	Weibull AFT	0.0909(0.0029)	0.4071(0.0581)	0.3882(0.0657)	3.336(0.234)
		Cox PH	0.0934(0.0005)	0.4018(0.0284)	0.3849(0.0305)	2.991(0.081)
		Ours	0.0942(0.0037)	**0.4003(0.0531)**	**0.3838(0.0588)**	2.861(0.076)