*Eclipta Journals*



RESEARCH

# Decision-Calibrated Multimodal Sequence Models for Rare-Event Forecasting in Heterogeneous Streams

Gerald Navarro,[1] Marvin Tolentino,[2] and Ramon Villareal[3]

[1]Department of Information Technology, Mindoro State University, Victoria Campus, Alcate Road, Victoria 5205, Philippines
[2]College of Computer Studies, Eastern Samar State University, Arteche Boulevard, Borongan City 6800, Philippines
[3]Department of Computer Engineering, Palawan State University, Rizal Avenue Extension, Puerto Princesa City 5300, Philippines

**Abstract**

Data streams in safety-critical environments increasingly combine irregular multivariate telemetry with sparse, high-level textual context. These heterogeneous signals are temporally misaligned, partially observed, and generated under nonstationary measurement policies, making rare-event forecasting a difficult learning problem despite the apparent abundance of logged data. Modern deep sequence models can rank future adverse events well, yet ranking alone is insufficient when downstream systems must trigger interventions under resource constraints and asymmetric costs. This paper develops a computer-science–focused methodology for decision-calibrated multimodal forecasting that treats reliability as a first-class design target rather than a post-hoc diagnostic. We formalize horizon-based risk prediction under censoring and repeated sampling, define modality-aware representations that degrade gracefully under missing channels, and derive training and calibration procedures that preserve discrimination while stabilizing operating thresholds. We further connect calibration to decision-theoretic optimization, showing how miscalibration interacts with class imbalance, prevalence drift, and selection bias induced by observation masks. The approach integrates gated fusion, uncertainty-aware objectives, and selective prediction to reduce false-trigger rates without sacrificing early-warning lead time. We outline evaluation protocols that separate improvements due to information gain from those due to base-rate stratification and propose drift monitors that are compatible with delayed labels. The resulting framework provides a principled path from multimodal encoders to actionable risk scores suitable for streaming deployment, emphasizing reproducible abstractions, provable properties where possible, and auditability under distribution shift.

## 1. Introduction

Rare-event forecasting in heterogeneous streams occupies an uncomfortable middle ground between classical signal processing and modern representation learning [1]. On one hand, the data are structured enough to admit formal dynamical descriptions, with measurable trajectories and identifiable constraints on temporal smoothness, sampling latency, and sensor noise. On the other hand, the streams are produced by complex socio-technical processes in which what is measured, when it is measured, and how it is described in text are all entangled with latent state and with operational decisions. As a result, the learning system does not observe a stationary stochastic process; it observes an interaction between latent dynamics and an observation policy that changes with context, workload, and perceived risk.

In many deployments, the model's output is not merely a score to be inspected after the fact. It is a control signal that either triggers an automated action, generates a ranked list for a human operator, or gates downstream computational resources. In each of these cases, the numerical meaning of the predicted probability matters because it determines thresholds, queue sizes, and cost tradeoffs. Even when the ultimate decision is ordinal, such as selecting the top-$k$ cases for review, stable probability estimates are valuable because they yield predictable alert volumes under drift and enable principled policies for abstention, escalation, and model monitoring. Consequently, a modeling pipeline that focuses solely on discrimination metrics risks producing a system that performs well in offline comparisons yet behaves erratically when embedded in a decision loop [2].

Multimodality exacerbates these issues. A high-frequency numeric stream provides granular dynamics but can be sparse in coverage due to irregular sampling and missingness. A textual stream supplies compressed semantics and human interpretations, but it is temporally coarse, often delayed, and frequently absent. Naively concatenating both channels can yield an overconfident model that relies on artifacts, such as

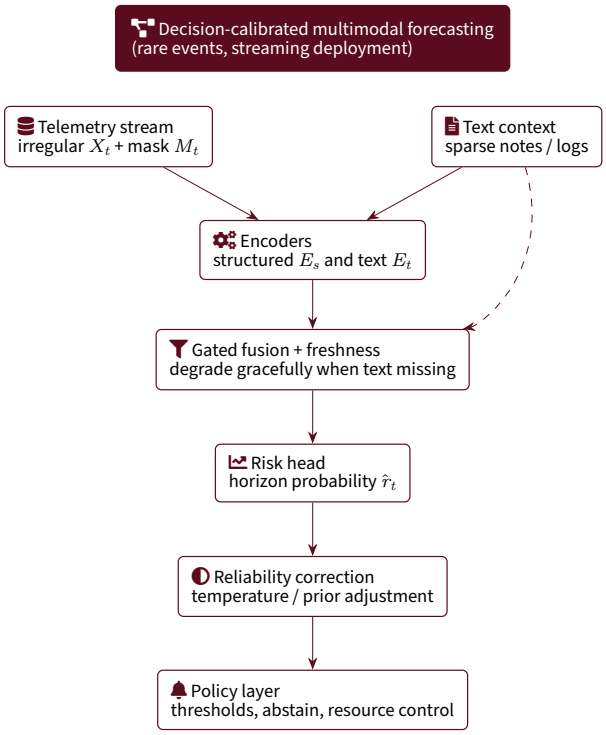



Figure 1. End-to-end pipeline. Irregular multivariate telemetry and sparse text are encoded separately, fused with freshness-aware gating, mapped to horizon risk, then post-processed for numerical reliability before entering a decision policy that controls alert volume and intervention cost.

the presence of text rather than its content, or measurement frequency rather than underlying state. A reliable multimodal model must therefore learn a representation in which each modality is used conditionally, with explicit mechanisms that detect staleness, missingness, and weak alignment, and that limit negative transfer when a modality is uninformative.

The central thesis of this paper is that decision-calibrated forecasting can be framed as a constrained optimization problem in which discrimination, calibration, and stability are jointly controlled by architectural inductive biases and by post-hoc reliability corrections. Instead of viewing calibration as a minor adjustment after training, we treat it as a structural requirement that shapes the choice of losses, sampling strategies, and fusion operators. This yields a design lens that is closer to systems engineering than to leaderboard optimization: the goal is not to maximize a single metric, but to bound failure modes that are likely to matter in real streaming deployments [3].

The paper proceeds by formalizing horizon-based forecasting with repeated predictions under censoring, making explicit the statistical dependence across time and the confounding role of observation masks. It then develops multimodal representation and alignment strategies that explicitly incorporate modality freshness and that allow the model to reduce to a unimodal predictor in the absence of text without distributional shock. The subsequent sections connect calibration to decision theory and provide mathematical characterizations of temperature scaling, prior-shift correction, and selective prediction under uncertainty constraints. We then address evaluation under shift, including stratified analyses that separate base-rate effects from genuine information gains, and propose monitoring signals that are compatible with delayed outcome labels. The paper closes with practical considerations for streaming inference, including computational complexity, caching strategies for text embeddings, and reproducible logging that supports retrospective audit.

Throughout, the emphasis is deliberately technical and domain-agnostic. Although the motivating examples resemble clinical prediction problems, the framework applies equally to industrial telemetry with maintenance logs, cybersecurity with analyst notes, content moderation with moderator rationales, and autonomous systems with diagnostic text. The unifying structure is a multimodal stream with irregular numeric measurements and sparse text that encodes high-level latent state, coupled to rare adverse events that must be forecasted within a finite horizon. In such settings, reliability is not optional, and the cost of miscalibration is not merely theoretical; it manifests as unstable alert rates, brittle thresholds, and untrustworthy numerical outputs that undermine system integration [4].

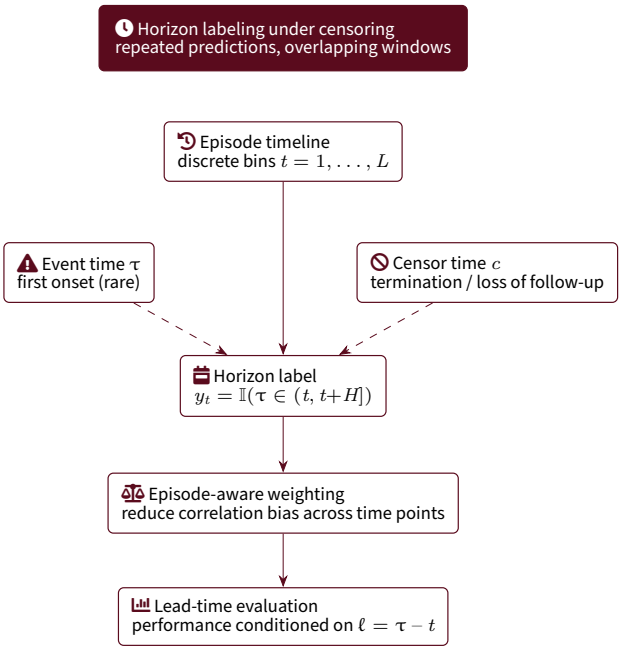



Figure 2. Repeated sampling with censoring. Horizon labels overlap heavily across adjacent times, and censoring truncates valid windows. Training and evaluation benefit from episode-aware sampling/weighting and lead-time–conditioned analyses that reveal whether signal appears early or only immediately before the event.

Table 1. Notation for horizon-based risk forecasting

| Symbol | Meaning | Domain |
|---|---|---|
| $X_{i,t}$ | Look-back window of structured features | $\mathbb{R}^{T\times d}$ |
| $M_{i,t}$ | Observation mask aligned with $X_{i,t}$ | $\{0,1\}^{T\times d}$ |
| $u_{i,t}$ | Text embedding available at time $t$ | $\mathbb{R}^p \cup \{\varnothing\}$ |
| $s_i$ | Static covariates for episode $i$ | $\mathbb{R}^q$ |
| $\tau_i$ | Event time (first occurrence) | $\{1, \ldots, L_i\} \cup \{\infty\}$ |
| $c_i$ | Censoring time | $\{1, \ldots, L_i\}$ |

## 2. Formal Risk Forecasting Under Censoring and Repeated Sampling

Consider an episode indexed by $i$ observed over discrete time steps $t \in \{1, \ldots, L_i\}$, where each step represents a fixed bin width in the deployed system. Let $X_{i,t} \in \mathbb{R}^{T\times d}$ denote a look-back window of structured features, $M_{i,t} \in \{0,1\}^{T\times d}$ the corresponding observation mask, $u_{i,t} \in \mathbb{R}^p \cup \{\varnothing\}$ a text embedding available at prediction time, and $s_i \in \mathbb{R}^q$ static covariates. The learning objective is to estimate the probability that a target event occurs within a horizon $H$ after time $t$, with repeated predictions over time.

A convenient abstraction introduces an event time $\tau_i$ measured in discrete steps, representing the first time the event initiates, and a censoring time $c_i$ representing episode termination or loss of follow-up. The observed terminal time is $\tilde{\tau}_i = \min(\tau_i, c_i)$ with an indicator $\delta_i = \mathbb{I}(\tau_i \le c_i)$. In horizon-based forecasting, the label is defined at each prediction time $t$ as $y_{i,t} = \mathbb{I}(\tau_i \in (t, t+H])$ with the convention that $y_{i,t} = 0$ if $t \ge \tilde{\tau}_i$ or if the event has already occurred. This creates an overlapping-label structure: adjacent prediction times share much of the same future window, implying strong dependence among labels within an episode.

The dependence has practical consequences. Standard i.i.d. generalization arguments become loose because the effective sample size is closer to the number of episodes than the number of time points. Moreover, the model can achieve deceptively good performance by learning signatures of imminent events that appear only near the end of the horizon, rather than true early-warning patterns. This can be analyzed by introducing a lead-time variable $\ell = \tau_i - t$ and evalu-

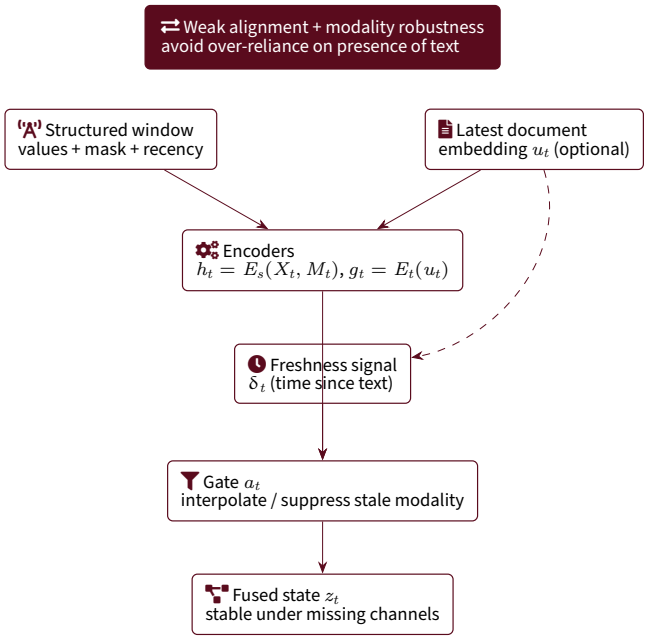



Figure 3. Modality-aware fusion. A freshness feature and a learned gate constrain cross-modal influence, enabling the model to default smoothly to the structured-only pathway when text is absent or stale, while still exploiting informative notes when available.

ating performance conditional on $\ell$. A model that is useful for early warning should retain discriminative power at larger $\ell$ rather than collapsing to near-zero signal until shortly before the event [5].

A risk predictor is a function $f_\theta$ that maps the observed history to a probability,

$$\begin{aligned} \hat{r}_{i,t} &= f_\theta(X_{i,t}, M_{i,t}, u_{i,t}, s_i) \\ \hat{r}_{i,t} &\approx \mathbb{P}(y_{i,t} = 1 \mid X_{i,t}, M_{i,t}, u_{i,t}, s_i) \end{aligned} \tag{1}$$

The conditioning on $M_{i,t}$ is essential in irregular streams because the mask is informative about the observation policy. Even if one imputes missing values, the mask carries signal about whether the value was observed, which may correlate with latent state and with future events. Treating the mask as exogenous can lead the predictor to exploit measurement intensity as a proxy for risk, yielding high discrimination in a particular environment but poor transportability when observation policies change.

A more explicit probabilistic model clarifies this entanglement. Let $z_{i,t}$ be a latent state evolving according to an unknown transition kernel, and let $o_{i,t}$ be an observation process that selects which components of $z_{i,t}$ are measured. Then $X_{i,t}$ is a noisy function of $(z_{i,t}, o_{i,t})$, while $M_{i,t}$ reflects $o_{i,t}$ directly. Text generation can be viewed as a process $d_{i,t}$ that produces documents from $(z_{i,t}, \eta_{i,t})$, where $\eta_{i,t}$ encodes social and operational context such as staffing, workflow, and reporting style. The target event time $\tau_i$ is a stopping time defined on the latent process. Under this view, the forecasting task is not simply estimating $\mathbb{P}(\tau_i \le t + H)$; it is estimating that probability under a particular observation and documentation policy. This matters because changes in policy induce distribution shift even if the underlying latent dynamics are stable.

One way to connect horizon-based labels to a hazard interpretation is to define a discrete-time hazard $h_{i,t} = \mathbb{P}(\tau_i = t \mid \tau_i \ge t, \mathcal{H}_{i,t})$ where $\mathcal{H}_{i,t}$ denotes the observed history at time $t$. The horizon risk can then be expressed via survival probabilities:

$$\begin{aligned} \hat{r}_{i,t} &= 1 - \prod_{k=1}^{H} \left(1 - \hat{h}_{i,t+k}\right) \\ \hat{h}_{i,t+k} &= g_\theta(\mathcal{H}_{i,t+k}) \end{aligned} \tag{2}$$

This decomposition suggests two modeling strategies. A direct strategy predicts $\hat{r}_{i,t}$ with a binary classifier trained on horizon labels. A hazard strategy predicts $\hat{h}_{i,t}$ at each time and converts hazards to horizon risks. The hazard approach yields a more structured temporal model and can support variable horizons, but it introduces additional dependence because hazards at adjacent times are correlated [6]. In practice, a direct horizon classifier can be simpler, while still permitting lead-time evaluation by conditioning on $\ell$.

Repeated sampling also creates multiple negatives that are nearly identical within an episode, which can dominate gradients and distort calibration. If an episode has no event, then all time points are negative, producing a large number of correlated negative samples. If an episode has an event, then early time points are negative and later time points become positive as

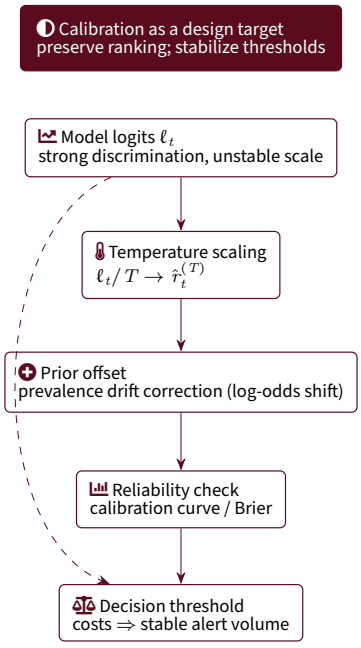



Figure 4. Reliability correction loop. Post-hoc monotone logit transformations (e.g., temperature plus optional prior adjustment) can repair probability scale without sacrificing discrimination, yielding thresholds that correspond to predictable operating points under class imbalance and drift.

the horizon window overlaps the event, creating structured label noise if the forecast horizon does not match the actionable intervention window. For these reasons, evaluation and training should consider episode-aware sampling schemes, such as subsampling time points or weighting time points to reduce correlation-driven bias.

Censoring further complicates evaluation. If the horizon extends beyond the censoring time, then the label is undefined because the event might occur after censoring. A conservative choice sets such labels to missing and excludes them from training and evaluation [7]. Another choice assumes no event after censoring, which can bias risk downward in episodes that terminate early. A principled approach is to exclude time points $t$ for which $t + H > c_i$, which ensures that the label window is fully observed. This reduces sample count but avoids a systematic label bias that can masquerade as calibration error.

The formalism above provides the scaffold for the remainder of the paper. The core challenge is to build $f_\theta$ in a way that is robust to the observation policy encoded by $M_{i,t}$, that exploits text when it adds information but does not rely on its presence, and that produces probabilities that are stable enough for decision integration. Achieving these properties requires both architectural constraints and explicit reliability procedures.

## 3. Multimodal Representation and Weak Temporal Alignment

A multimodal streaming predictor can be decomposed into modality-specific encoders and a fusion operator. The structured encoder processes irregular multivariate sequences, while the text encoder processes unstructured language into a vector embedding. The fusion operator combines them into a shared representation used by a classifier head. The core technical difficulty is that the modalities are not aligned: structured measurements may arrive hourly or faster, while text may arrive sporadically and refer to intervals rather than instants [8]. Moreover, text may be unavailable for a nontrivial fraction of episodes, making it an optional modality whose absence must not cause distributional discontinuities.

Let $E_s$ be a structured encoder and $E_t$ a text encoder. For prediction at time $t$, define

$$\begin{aligned} h_{i,t} &= E_s(X_{i,t}, M_{i,t}) \\ g_{i,t} &= E_t(u_{i,t}) \\ z_{i,t} &= \Phi(h_{i,t}, g_{i,t}, s_i) \\ \hat{r}_{i,t} &= \sigma(w^\top z_{i,t} + b) \end{aligned} \tag{3}$$

The encoders can be implemented in several ways. For structured sequences, recurrent networks, temporal convolutions, and attention-based architectures each offer different tradeoffs. Recurrent encoders are efficient for moderate window length $T$ and can incorporate missingness indicators directly. Attention-based encoders can model long-range dependencies and variable importance across time, but they can be expensive and can

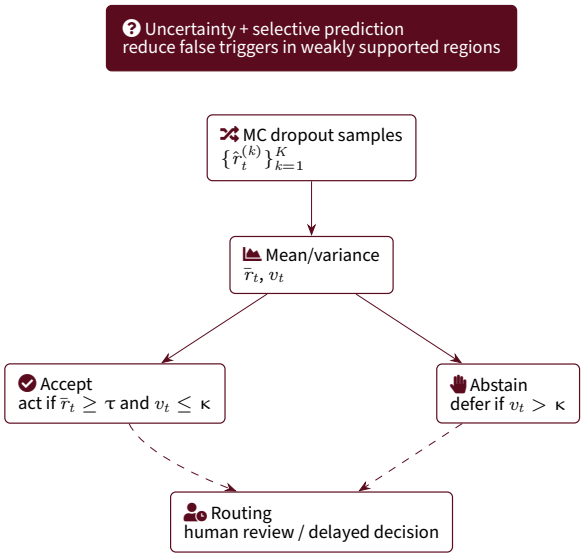



Figure 5. Selective prediction policy. A lightweight uncertainty proxy (variance across stochastic forward passes) enables abstention rules that trade coverage for lower risk, especially when modality patterns or missingness regimes differ from training support.

Table 2. Direct horizon classifier versus discrete-time hazard model

| Aspect | Horizon classifier | Hazard model |
|---|---|---|
| Target | $\mathbb{P}(\tau_i \in (t, t+H])$ | Hazard $h_{i,t} = \mathbb{P}(\tau_i = t \mid \tau_i \geq t)$ |
| Labels | Overlapping binary labels per $(i, t)$ | Event-time labels with survival constraints |
| Horizon handling | Fixed deployment horizon $H$ | Multiple horizons via survival curve composition |
| Modeling structure | Generic binary sequence classifier | Structured temporal model over $\{t+1, \ldots, t+H\}$ |
| Lead-time analysis | Post-hoc by conditioning on lead time $\ell$ | Natural via hazards over future steps |
| Complexity | Simple training and inference | Extra parameters and dependence across time |

overfit when the effective sample size is the number of episodes rather than the number of time points.

A missingness-aware structured representation is often built by augmenting each time step with both observed values and mask-derived features [9]. For each time step $\tau$ within the look-back window, define an augmented vector $\tilde{x}_\tau$ that concatenates normalized values, observation indicators, and a time-since-last-observed feature. This gives the encoder explicit access to the observation policy, allowing it to learn, for example, that a sudden increase in measurement frequency can be informative, while also allowing it to condition on actual values when they exist. A structured encoder that ignores $M_{i,t}$ will implicitly treat imputed values as real and may develop brittle dependencies on the imputation scheme.

Text embeddings are typically computed by a pretrained language model and projected into the same latent dimension as the structured embedding. When text is sparse, a common strategy is to use the most recent document embedding available at prediction time, resulting in a piecewise-constant $g_{i,t}$ that remains unchanged across many time steps. This can be beneficial because it provides stable context, but it can also introduce misleading invariances: the model may assume the same textual context is valid throughout the interval, even when the latent state changes rapidly. A more robust representation adds a freshness feature $\delta_{i,t}$ indicating time since the last document, enabling the fusion module to discount stale text. In addition, an indicator that captures whether any text exists prevents the model from interpreting a zero vector as meaningful content.

The fusion operator $\Phi$ is the linchpin of robustness. Pure concatenation of $h_{i,t}$ and $g_{i,t}$ is expressive but can be unstable under missing text, because the distribu-

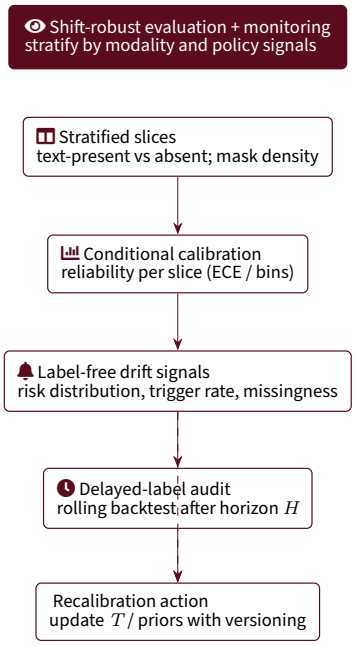



Figure 6. Evaluation and monitoring under drift. Robust reporting separates information gain from base-rate artifacts by stratifying on modality availability and observation masks, while deployment monitoring combines label-free distribution checks with delayed-label calibration audits that trigger conservative recalibration updates.

Table 3. Key components in heterogeneous streams

| Component | Signal type | Strengths | Vulnerabilities |
|---|---|---|---|
| Structured numeric stream | High-frequency telemetry | Fine-grained dynamics, early changes | Irregular sampling, artefactual missingness |
| Text notes | Sparse human-authored summaries | Rich semantics and latent state | Coarse, delayed, presence tied to severity |
| Static covariates $s_i$ | Episode-level descriptors | Stable baseline risk stratification | Encode selection effects and demographics |
| Observation mask $M_{i,t}$ | Measurement pattern over time | Exposes policy and workload | Easy to overuse as proxy for acuity |
| Lead time $\ell$ | Time to event | Makes early-warning utility explicit | Shrinks near censoring and episode end |

tion of concatenated vectors differs drastically between the present-text and absent-text regimes. A gating-based fusion operator introduces an inductive bias that explicitly interpolates between modalities. Let a gate $a_{i,t} \in (0,1)^m$ be computed as a function of both embeddings and freshness:

$$\begin{aligned} a_{i,t} &= \sigma\big(G\,[h_{i,t}; g_{i,t}; s_i; \delta_{i,t}] + c\big) \\ z_{i,t} &= a_{i,t} \odot h_{i,t} + (1 - a_{i,t}) \odot g_{i,t} \end{aligned} \tag{4}$$

This representation has several desirable properties. When text is absent, $g_{i,t}$ can be set to zero and $\delta_{i,t}$ can be set to a sentinel value, allowing the gate to learn $a_{i,t} \approx 1$ and reduce the model to a structured-only predictor. When text is present but stale, the gate can learn to emphasize $h_{i,t}$, while still permitting $g_{i,t}$ to contribute long-term context such as chronic conditions or persistent constraints. When text is fresh and informative, the gate can shift weight toward $g_{i,t}$ or toward cross-modal interactions.

Cross-modal attention provides a richer interaction mechanism. Instead of interpolating between two fixed vectors, cross-attention allows text to reweight structured temporal features [10]. If $H_{i,t} \in \mathbb{R}^{T \times m}$ denotes the matrix of structured hidden states over the window and $g_{i,t} \in \mathbb{R}^m$ the projected text vector, then a single-head attention can compute a text-conditioned summary of the structured trajectory:

$$\begin{aligned} q_{i,t} &= W_q g_{i,t} \\ K_{i,t} &= H_{i,t} W_k \quad V_{i,t} = H_{i,t} W_v \\ \beta_{i,t} &= \mathrm{softmax}\left(\frac{q_{i,t}^\top K_{i,t}^\top}{\sqrt{m}}\right) \\ h^\star_{i,t} &= \beta_{i,t} V_{i,t} \end{aligned} \tag{5}$$

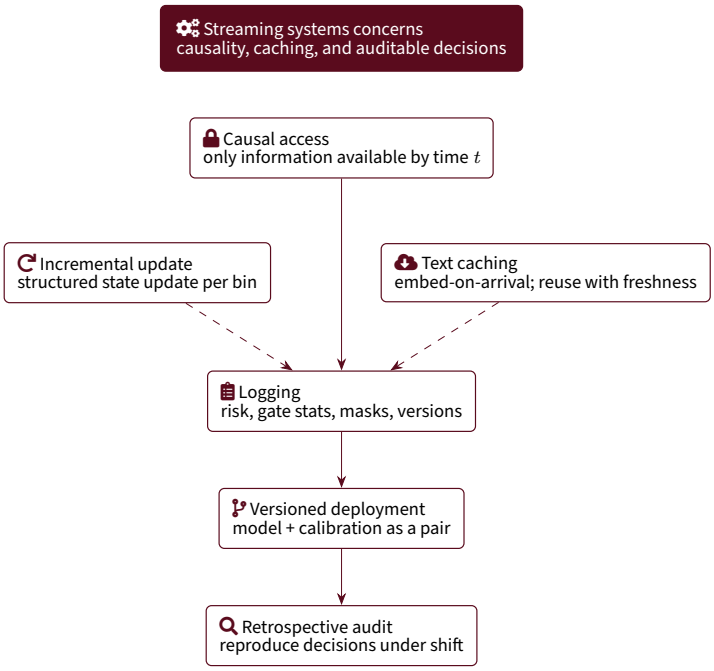



Figure 7. Deployment engineering view. Streaming inference must respect temporal validity while keeping compute bounded through incremental updates and cached text embeddings; reliable integration requires structured logging and strict versioning of both the base model and its calibration mapping for audit and governance.

Table 4. Fusion designs for weakly aligned modalities

| Fusion design | Description | Advantages | Failure modes |
|---|---|---|---|
| Concatenation | Stack $h_{i,t}$ and $g_{i,t}$ then MLP | Maximal flexibility and capacity | Brittle under missing text or shift in text policy |
| Gated interpolation | Learn gate $a_{i,t}$ over modalities | Graceful degradation when text absent or stale | Gate can collapse to single modality |
| Cross-modal attention | Text queries structured history | Aligns phrases to temporal segments | Expensive; may overfit rare patterns |
| Modality dropout | Randomly drop text during training | Robustness to missing channels | If too strong, underuses informative text |
| Text-only branch | Separate text-based predictor | Clarifies marginal value of textual context | Ignores numeric early-warning signals |

The resulting $h^{\star}_{i,t}$ can then be fused with $g_{i,t}$ and $s_i$. This design separates two roles: text selects relevant temporal segments via $\beta_{i,t}$, while the structured trajectory contributes fine-grained dynamics. In weakly aligned settings, this can reduce the tendency for text to dominate decisions purely through presence, because it must exert influence through a learned selection over structured time steps.

Weak alignment also suggests training-time augmentations that simulate temporal uncertainty. For example, one can randomly jitter the selected note within a plausible window, or randomly drop text for a fraction of training samples even when it exists, forcing the model to remain competent without text. Such stochastic modality dropout is a regularization technique that can improve robustness under missingness at deployment. The key is to perform dropout in a way that matches expected deployment conditions; otherwise the model may become overly conservative and underuse text even when it is informative.

Finally, representation learning should respect the asymmetry of information density. Structured sequences contain many time steps, each with multiple features, while text is often a single vector. The model should avoid allocating disproportionate capacity to the text projection head relative to the structured encoder, because high-capacity text heads can memorize idiosyncratic embeddings correlated with outcomes through spurious pathways [11]. In practice, this means limiting the depth of the fusion network, using dropout on the text projection, and monitoring performance separately for text-present and text-absent strata. These

Table 5. Calibration and reliability corrections

| Method | Parameters | Preserves ranking? | Typical use |
|---|---|---|---|
| None (raw logits) | – | Yes | Baseline; often overconfident under heavy imbalance |
| Temperature scaling | Single scalar $T$ | Yes | Low-variance fix for global miscalibration |
| Isotonic regression | Nonparametric step function | Not guaranteed | Flexible but data hungry for rare events |
| Platt scaling | Affine logit transform | Yes | Lightweight sigmoid recalibration |
| Prior-shift offset | Logit shift $\Delta$ from new prevalence | Yes | Correct pure prior drift under stable conditionals |
| Per-stratum temperature | One $T$ per evaluation slice | Within slice | Align calibration across modality regimes |

Table 6. Uncertainty mechanisms for selective prediction

| Mechanism | Uncertainty type | Strengths | Limitations |
|---|---|---|---|
| MC dropout variance | Epistemic over weights | Uses existing network structure | Multiple forward passes increase latency |
| Predictive entropy | Mixed aleatoric and epistemic | Cheap scalar from $\hat{r}_{i,t}$ | Conflates class imbalance and ambiguity |
| Reliability head | Empirical error probability | Learns data-dependent reliability | Needs dedicated labeled calibration set |
| Conformal prediction sets | Set size or abstain region | Finite-sample coverage guarantees | Exchangeability violated under temporal drift |
| Entropy regularizer | Encouraged uncertainty without text | Prevents unwarranted confidence when $u_{i,t} = \varnothing$ | May suppress true high-risk cases |
| Heuristic OOD scores | Distance in representation space | Low overhead, label free | Hard to map to explicit decision costs |

design choices set the stage for the reliability focus of the next sections, where calibration is treated not as an optional add-on but as an integral property of the deployed predictor.

## 4. Calibration Theory, Temperature Scaling, and Reliability Correction

A calibrated predictor is one whose probabilities match empirical frequencies in the long run. In a binary forecasting task, calibration can be expressed as $\mathbb{P}(y = 1 \mid \hat{r} = p) = p$ for all $p$ in the support. In finite samples, calibration is assessed through binning-based statistics or through proper scoring rules such as the Brier score and negative log-likelihood. Calibration matters because decisions are typically based on thresholds or on cost-sensitive policies that assume numerical probabilities are meaningful. When calibration is poor, the same threshold can generate wildly different alert volumes across environments, and cost-sensitive optimization becomes unstable.

Neural sequence models often exhibit miscalibration, particularly when trained under heavy class imbalance and with losses that emphasize difficult positives. This is not merely an empirical annoyance; it is a predictable outcome of optimization dynamics and of the mismatch between the training objective and the desired decision environment. When training uses class reweighting or focal loss, the effective likelihood being optimized differs from the true likelihood under the deployment class prior [12]. Even when cross-entropy is used, strong regularization, early stopping, and correlated samples can yield overconfident logits.

Temperature scaling is a simple and widely applicable post-hoc correction that rescales logits by a single scalar $T > 0$ to reduce overconfidence while preserving ranking. If a model produces a logit $\ell_{i,t}$, then the

Table 7. Shift-aware evaluation strata

| Stratification variable | Example slices | Purpose |
|---|---|---|
| Modality availability | Text-present vs text-absent episodes | Detect over-reliance on text presence |
| Mask density | High vs low measurement frequency | Separate acuity from information volume |
| Time period | Pre- vs post-policy or workflow change | Quantify temporal distribution shift |
| Entity group | Site, region, device family, or cohort | Assess transportability across entities |
| Lead time $\ell$ | Early vs late prediction windows | Check usefulness as an early-warning system |

Table 8. Drift monitoring signals in streaming deployment

| Monitor | Signal | Needs labels? | Typical trigger |
|---|---|---|---|
| Risk histogram | Distribution of $\hat{r}_{i,t}$ | No | Shift in mass at high or low risk |
| Alert rate | Fraction above action threshold | No | Sustained change in alert volume |
| Mask statistics | Mean density and pattern counts of $M_{i,t}$ | No | Abrupt shifts in measurement policy |
| Text freshness | Time since last available $u_{i,t}$ | No | New documentation or note-taking workflow |
| Rolling calibration | Bin-wise observed vs predicted rates | Yes (delayed) | Systematic over- or underprediction |
| Prevalence estimate | Recent positive rate per window | Yes (delayed) | Class-prior drift requiring adjustment |

temperature-scaled probability is

$$\hat{r}_{i,t}^{(T)} = \sigma\left(\frac{\ell_{i,t}}{T}\right)$$
$$\ell_{i,t} = \log\left(\frac{\hat{r}_{i,t}}{1-\hat{r}_{i,t}}\right) \tag{6}$$

The parameter $T$ is typically fit by minimizing negative log-likelihood on a held-out calibration set drawn from the deployment distribution. The key theoretical property is that scaling by $T$ is a monotone transformation of logits, so any rank-based metric such as AUROC is unchanged, while calibration can improve substantially if miscalibration is primarily due to overconfident logits rather than to systematic ranking errors.

In practice, the effectiveness of temperature scaling hinges on how well a single scalar can correct the model's logit distribution. For many deep classifiers, miscalibration is approximately uniform across score ranges, especially when the miscalibration arises from a global margin increase induced by training dynamics. In such cases, a scalar temperature can yield large improvements in ECE and Brier score without affecting discrimination. This phenomenon is consistent with the observation that a strongly discriminative model can still be numerically unreliable if its logit scale is mismatched to the empirical prevalence [13].

A concrete illustration of this reliability gap arises in multimodal rare-event forecasting settings where focal loss is used to cope with severe imbalance. Sadanandan (2025) reports that applying temperature scaling with an optimized temperature $T = 0.26$ reduced expected calibration error from 0.2129 to 0.0240 and reduced Brier score from 0.0760 to 0.0303 while leaving discrimination metrics unchanged [14], highlighting that post-hoc scaling can repair probability reliability even when the underlying ranking is already strong. The technical implication is that calibration should be treated as a distinct axis of model quality and addressed with explicit procedures rather than inferred from discrimination performance.

Temperature scaling can be derived from a decision-theoretic perspective. Suppose the downstream policy triggers an action when $\hat{r} \geq \tau$, where $\tau$ is chosen based on costs $C_{\mathrm{FP}}$ and $C_{\mathrm{FN}}$ under a simple expected-cost model. If probabilities are miscalibrated, then the effective cost ratio implied by $\tau$ differs from the intended one, leading to either excessive false triggers or excessive misses. If miscalibration is mostly a logit-scale issue, then temperature scaling rescales $\ell$ so that the

Table 9. Systems constraints and mitigations for streaming inference

| Constraint | Failure mode | Mitigation | Implementation note |
|---|---|---|---|
| Causal access only | Leakage from future into past | Enforce strict timestamp filters | Test with artificial time shifts |
| Latency budget | Missed real-time deadlines | Incremental state updates; slim heads | Profile worst-case sequence lengths |
| Expensive text encoding | Throughput collapse on note bursts | Precompute and cache embeddings | Use frozen encoder keyed by document time |
| Versioned calibration | Mixing models and temperatures | Jointly version model and calibrator | Log model and calibration IDs with outputs |
| Limited storage | Incomplete audit trail | Log compact summaries and hashes | Preserve enough for drift and fairness analysis |
| Privacy constraints | Inability to share raw logs | Aggregate monitoring statistics | Design privacy-aware audit pathways |

Table 10. Pre-deployment checks for decision-calibrated streaming models

| Check | Metric | Expected range | Degradation signal |
|---|---|---|---|
| Modality-drop test | AUROC and AUPRC without text | Moderate but controlled drop | Collapse when $u_{i,t} = \varnothing$ |
| Mask-perturbation test | Discrimination under extra missingness | Small, smooth degradation | Highly unstable thresholds |
| Stratified calibration | ECE per slice | Similar across strata | One slice much worse than global |
| Threshold stability | Alert rate at fixed cutoff | Within design band on recent data | Large swings across time windows |
| Reproducibility audit | Hash match of inputs and outputs | Deterministic replay | Non-deterministic or drifting predictions |

implied cost ratio matches observed frequencies more closely. From this viewpoint, calibration is not merely about matching frequencies; it is about ensuring that thresholds correspond to stable operating points.

More complex reliability corrections go beyond a single scalar. One can fit isotonic regression on logits to produce a nonparametric calibration curve, or fit a piecewise linear transformation that better matches bin-wise frequencies. These methods can overfit when positives are rare, because calibration bins at high predicted probabilities may contain very few samples [15]. A scalar temperature has the advantage of low variance. In rare-event streaming contexts, low-variance corrections are often preferable because they maintain stability across time, which is itself a reliability requirement.

Calibration under repeated sampling requires additional care. If each episode contributes many time points, and if calibration is evaluated at the time-point level, then episodes with long durations can dominate the calibration curve. This is undesirable because long episodes may correspond to different regimes of observation policy. A more balanced approach calibrates at the episode level by selecting representative time points, or uses cluster-weighted calibration that assigns equal weight to each episode. Similarly, if calibration is intended for a policy that triggers once per episode, then time-point calibration may be a mismatched target. Reliability should be evaluated at the decision granularity of the deployed system [16].

Calibration must also confront prevalence drift. If the positive rate changes, then a perfectly calibrated model under the old prevalence becomes miscalibrated under the new prevalence, even if class-conditional feature distributions remain stable. A prior-shift correction can adjust logits by a constant offset:

$$\begin{aligned} \ell'_{i,t} &= \ell_{i,t} + \Delta \\ \Delta &= \log\left(\frac{\pi_{\text{new}}}{1-\pi_{\text{new}}}\right) \\ &\quad - \log\left(\frac{\pi_{\text{old}}}{1-\pi_{\text{old}}}\right) \end{aligned} \tag{7}$$

This adjustment is valid under strict prior shift assumptions. In multimodal streams, prior shift can be confounded with observation-policy shift, so the correction should be applied cautiously and validated with delayed labels. Still, the form of the correction highlights a central theme: reliability is not a single property of the model; it is a property of the model plus

the environment, and it must be maintained through explicit monitoring and adjustment mechanisms.

## 5. Uncertainty Quantification and Selective Prediction

Even a calibrated model can be unreliable on out-of-distribution inputs, on rare subpopulations, or under missing modality conditions that were underrepresented in training. Uncertainty quantification aims to detect such cases and to express predictive ambiguity in a way that supports safer decision policies. In streaming deployments, a practical uncertainty strategy is selective prediction, where the model abstains or defers when uncertainty exceeds a threshold, thereby controlling risk at the expense of coverage [17].

Uncertainty in deep models has multiple sources. Aleatoric uncertainty arises from intrinsic noise and label ambiguity, such as overlapping feature distributions between positives and negatives. Epistemic uncertainty arises from limited data support and model misspecification, such as novel combinations of missingness patterns and text contexts. A deployed system benefits most from epistemic uncertainty signals because they can indicate when the model is likely to be wrong due to insufficient training coverage.

A tractable approach to epistemic uncertainty is Monte Carlo dropout. By applying dropout at inference time and sampling multiple stochastic forward passes, one can approximate a posterior predictive distribution. Let $\hat{r}_{i,t}^{(k)}$ be the prediction from sample $k$ and define the mean and variance:

$$\begin{aligned} \bar{r}_{i,t} &= \frac{1}{K}\sum_{k=1}^{K} \hat{r}_{i,t}^{(k)} \\ v_{i,t} &= \frac{1}{K-1}\sum_{k=1}^{K} \left(\hat{r}_{i,t}^{(k)} - \bar{r}_{i,t}\right)^2 \end{aligned} \tag{8}$$

The variance $v_{i,t}$ can be used as an uncertainty proxy. In a selective prediction policy, one can trigger actions only when $\bar{r}_{i,t}$ exceeds a risk threshold and $v_{i,t}$ is below an uncertainty threshold, or equivalently abstain when uncertainty is high. This reduces false triggers in poorly supported regions but requires careful tuning to avoid systematically abstaining on the very cases that matter.

Selective prediction can be framed as a constrained optimization [18]. Define a decision $d_{i,t} \in \{0, 1, \perp\}$, where $\perp$ denotes abstention. Suppose the system aims to minimize expected cost subject to a minimum coverage constraint. Let $C_{\text{act}}$ be the cost of acting when no event occurs, $C_{\text{miss}}$ the cost of missing an event, and $C_{\text{def}}$ the cost of deferring, which may represent routing to a human operator. Then a policy that uses both risk and uncertainty can be described by thresholds $(\tau, \kappa)$:

$$\begin{aligned} d_{i,t} &= 1 \quad \text{if } \bar{r}_{i,t} \geq \tau \ \wedge \ v_{i,t} \leq \kappa \\ d_{i,t} &= \perp \quad \text{if } v_{i,t} > \kappa \\ d_{i,t} &= 0 \quad \text{otherwise} \end{aligned} \tag{9}$$

The effectiveness of this policy depends on whether $v_{i,t}$ correlates with error. This correlation is not guaranteed, especially under severe class imbalance, where uncertainty estimates may be dominated by the model's uncertainty about rare positives rather than by distribution shift. Therefore, uncertainty policies should be evaluated with metrics that measure risk conditional on acceptance, such as selective risk curves.

An alternative approach constructs prediction sets with guaranteed coverage using conformal prediction. In binary classification, conformal methods can output a set $\{0\}$, $\{1\}$, or $\{0, 1\}$, where $\{0, 1\}$ corresponds to abstention. The method calibrates a nonconformity score on a calibration set and chooses a threshold that guarantees that the true label lies in the set with probability at least $1 - \alpha$ under exchangeability assumptions. In streaming contexts, strict exchangeability is violated by temporal dependence and drift, but conformal ideas remain useful as a conceptual tool for designing abstention rules with explicit coverage targets.

Uncertainty also interacts with multimodal fusion [19]. When text is absent, the model's epistemic uncertainty may increase because a modality that could resolve ambiguity is missing. A fusion network can expose a modality-availability feature and learn to increase predictive entropy in such cases, effectively representing uncertainty about latent context. This can be encouraged by training objectives that penalize overconfident predictions when modality inputs are missing. For example, one can add an entropy regularizer on predictions from text-absent samples to discourage extreme probabilities unless supported by structured evidence:

$$\begin{aligned} \mathcal{L}_{\text{ent}} = -\lambda \sum_{i,t} & \mathbb{I}(u_{i,t} = \varnothing) \\ & \cdot \left(\hat{r}_{i,t} \log \hat{r}_{i,t} \right. \\ & \left. + (1 - \hat{r}_{i,t}) \log(1 - \hat{r}_{i,t})\right) \end{aligned} \tag{10}$$

Such a regularizer must be used cautiously because it can suppress true positives. The goal is not to force uncertainty, but to prevent the model from using missingness as a spurious certainty signal.

A complementary strategy is to model prediction as a two-stage process. The first stage produces a calibrated risk score. The second stage predicts a reliability score indicating whether the risk score is likely to be accurate, based on features such as mask density, text freshness, and embedding norms [20]. This creates a meta-model that can learn empirical reliability patterns even when uncertainty proxies like dropout

variance fail. The reliability model can be trained on held-out data by labeling each prediction as correct or incorrect at a chosen threshold and learning to predict correctness. While this introduces a dependence on the chosen threshold, it aligns with deployment where a threshold is often fixed.

Ultimately, uncertainty quantification is valuable only insofar as it improves system-level outcomes. It should be evaluated not as an aesthetic property but as a control primitive that trades off coverage against error under realistic decision costs. In multimodal rare-event forecasting, the most practical objective is to prevent pathological behavior under missing modality and shift, while preserving sufficient sensitivity to provide early warning. Selective prediction offers a mechanism to formalize this tradeoff and to integrate uncertainty into the architecture and evaluation pipeline.

## 6. Shift-Robust Evaluation and Drift Monitoring

Offline evaluation of rare-event predictors is deceptively simple: compute AUROC and AUPRC on a held-out set [21]. Yet in heterogeneous streams, distribution shift is the norm, not the exception. Shifts occur across time as policies change, across subpopulations as data collection differs, and across modality regimes as text availability varies. A shift-robust evaluation protocol must therefore go beyond aggregate metrics and analyze performance conditional on shift variables that are likely to change at deployment.

A first principle is that splitting should be performed at the episode or entity level to prevent leakage through repeated sampling. If the same entity appears in both training and test, the model can exploit stable identity-related features, including subtle linguistic patterns in text embeddings, to achieve inflated performance. Entity-level splitting also better approximates deployment, where the model is applied to new entities or new episodes.

A second principle is to stratify evaluation by modality availability and by missingness. In many systems, the presence of text correlates with the severity of the underlying state or with the type of episode, so improvements in multimodal models can be partly attributable to base-rate differences rather than to genuine information gain. Stratified metrics can reveal whether the model performs well because it uses text semantics or because it uses text presence as a proxy for a high-prevalence subgroup [22]. Similarly, stratification by mask density can reveal whether high missingness corresponds to a different acuity regime with a different positive rate, which can mechanically inflate AUPRC. These effects are not merely statistical curiosities; they determine whether a model will generalize when the measurement policy changes.

Shift-robust evaluation also requires calibration analyses that are conditional, not only marginal. A model may be globally calibrated but miscalibrated in a critical subgroup, such as episodes without text or episodes with sparse measurements. Conditional calibration can be assessed by computing calibration curves within strata and comparing ECE across strata. Because positives are rare, binning must be chosen carefully to avoid empty bins. Adaptive binning based on quantiles of predicted probabilities can improve stability, though it changes interpretability because bins vary across strata.

Beyond stratification, one can perform stress tests that simulate plausible shifts [23]. A modality-drop stress test randomly removes text from a fraction of text-present samples at evaluation time, measuring how performance degrades. A mask-perturbation stress test artificially increases missingness by dropping a random subset of structured values while keeping the mask consistent with the drop, approximating a scenario where measurement frequency decreases. A freshness stress test delays text embeddings by simulating staleness. These tests do not perfectly mimic real shifts, but they can expose brittle dependencies and provide comparative evidence about robustness.

Drift monitoring in deployment is constrained by label delay. If the forecast horizon is $H$, then ground-truth outcomes for predictions made at time $t$ are not fully observed until time $t+H$. This creates a monitoring lag. Therefore, monitoring should combine label-free signals with delayed label-based checks. Label-free signals include the marginal distribution of predicted risks, the rate of high-risk triggers, the distribution of mask densities, and the proportion of text-present predictions [24]. If these distributions change abruptly, the model may be operating in a new regime. Such signals are not definitive, but they can trigger closer inspection.

A label-based monitor can be implemented as a rolling calibration check with delayed reconciliation. For predictions made in a window $[t_0, t_1]$, once outcomes are observed after $H$ steps, compute the empirical event rate among predictions within risk bins and compare it to the predicted average. The lag implies that the monitor always operates on historical predictions, but it can still detect persistent drift. In settings where the environment changes gradually, this is sufficient to trigger recalibration or retraining.

Drift monitors should also account for changes in class prior. If the positive rate increases due to operational changes, the model's probabilities become miscalibrated even if the ranking remains good [25]. A practical monitor estimates the recent positive rate once labels arrive and compares it to the historical calibration prior. If the difference is large, a prior-shift adjustment or a recalibration procedure can be applied. However, prior estimation in rare events is noisy, so the monitor should aggregate over sufficiently long windows and incorporate uncertainty bands.

A deeper issue arises when drift is driven by observation-policy changes that also change the feature distribution conditional on the label. In that case, prior-shift correction is insufficient. One can detect such drift by comparing feature distributions, mask distributions, and embedding norms over time. For example, if the mean mask density changes while the predicted risk distribution remains constant, the model may be compensating by relying more heavily on missingness, which can be problematic if missingness becomes less informative in the new regime. Similarly, if text embedding norms change due to a new documentation template, the fusion module may misinterpret the embeddings, leading to unpredictable gating behavior.

A robust evaluation and monitoring strategy therefore treats shift variables as explicit objects [26]. Modality availability, mask density, and text freshness are not nuisances; they are state variables of the sociotechnical pipeline. By logging them, stratifying evaluation by them, and monitoring their distributions, the system can detect when the conditions under which calibration was established no longer hold. This is essential for any deployment that relies on numerical probabilities and fixed thresholds.

## 7. Systems Optimization for Streaming Inference and Reproducible Audit

The technical design of a multimodal predictor is inseparable from its systems implementation. Training-time conveniences that violate deployment constraints can yield a model that is impressive in offline evaluation but infeasible or unsafe in a streaming environment. Similarly, deployment without reproducible logging and audit trails undermines the ability to diagnose failures, recalibrate reliably, and demonstrate governance.

Streaming inference imposes strict causality: the model at time $t$ may use only information available up to $t$. In a look-back window formulation, this means all structured inputs must be drawn from $\{t-T+1, \ldots, t\}$, and text embeddings must correspond to documents available at or before $t$. If text embeddings are precomputed offline, caching must be keyed by document timestamp and retrieval must enforce the temporal constraint [27]. A common systems pitfall is to cache a single embedding per episode and reuse it without time checks, inadvertently allowing late documents to influence early predictions. Preventing this requires careful data engineering, not merely model design.

Computationally, the structured encoder often dominates inference if it processes a long window. However, in a streaming setting, one can update representations incrementally rather than recomputing the entire window at each time. For recurrent encoders, the hidden state can be updated with each new time step, reducing inference cost from $O(T)$ per step to $O(1)$ per step after initialization. For attention-based encoders, incremental updates are more complex because attention attends to the entire history, but one can restrict attention to a sliding window or use approximate attention mechanisms.

Text embeddings can be expensive if computed by a large language model. If the text encoder is frozen, embeddings can be computed when a document arrives and stored, then reused for subsequent predictions until a new document arrives. This reduces average cost and aligns with the piecewise-constant nature of text context [28]. The fusion module must then incorporate freshness so that the reuse does not imply equal relevance at all times.

The computational profile of a gated fusion model is modest. Let $m$ be the latent dimension. If the structured encoder outputs $h_{i,t} \in \mathbb{R}^m$ and the text projection outputs $g_{i,t} \in \mathbb{R}^m$, then the gate computation is a matrix-vector product with cost $O(m^2)$ if implemented densely. In practice, $m$ can be chosen to balance accuracy and throughput. For streaming systems with strict latency, one can reduce $m$, use low-rank projections, or quantize the fusion and classifier layers.

A reproducible audit trail requires logging at multiple layers. At minimum, the system should log the prediction time, the predicted probability, the threshold decision, modality availability indicators, mask density summaries, and text freshness. Logging raw inputs is often infeasible due to privacy or storage, but summary statistics and hashed identifiers can support drift detection [29]. For retrospective debugging, it is helpful to log intermediate quantities such as gate values or attention entropy, because they reveal whether the model relied on text or on structured data. Such logs also support fairness analyses by enabling stratification by demographic covariates, if those covariates are available and appropriately governed.

Reproducibility also concerns model versioning and calibration parameters. A calibrated deployment is a pair: a base model and a calibration mapping such as a temperature. If the base model is updated, the calibration mapping may no longer be valid. Therefore, the deployment should version the base model and the calibration mapping together and ensure that each prediction can be traced to a specific version. In addition, if prior-shift adjustments are applied dynamically, the estimated prior and the adjustment applied should be logged, because they affect numerical probabilities and thresholds.

Another systems concern is the interaction between retraining and monitoring. Frequent retraining can improve adaptation under drift, but it can also destabilize behavior if calibration is not re-established [30]. A disciplined approach is to decouple representation updates from calibration updates. The base model can be retrained or fine-tuned periodically, but calibration

should be recomputed on a recent calibration set that matches the intended environment. If calibration sets are too small due to rare positives, then calibration updates should be conservative, favoring low-variance methods such as temperature scaling.

Finally, the system should support counterfactual and robustness tests as part of a continuous integration pipeline. Before deploying a new model version, the system can run modality-drop and mask-perturbation tests, compute stratified calibration curves, and verify that alert volume at a fixed threshold remains within acceptable bounds on a recent validation slice. These checks are not perfect guarantees, but they operationalize reliability as a testable property rather than a hoped-for outcome.

In summary, streaming deployment requires more than a high-performing model. It requires causally valid data access, computational strategies that meet latency constraints, and logging that supports delayed-label monitoring and audit [31]. These systems constraints feed back into model design, particularly in multimodal settings where text processing and alignment introduce unique risks. A decision-calibrated framework must therefore be implemented as an end-to-end pipeline that preserves temporal validity, maintains calibration, and provides visibility into modality usage.

## 8. Conclusion

This paper developed a technical framework for decision-calibrated multimodal forecasting in heterogeneous streaming environments characterized by irregular numeric measurements, sparse textual context, severe class imbalance, and distribution shift. We formalized horizon-based risk prediction under censoring and repeated sampling, emphasizing the statistical dependence and observation-policy confounding that distinguish streaming logs from i.i.d. datasets. We then described modality-aware representation and weak alignment strategies that use gating and cross-modal attention to exploit text when informative while degrading gracefully when text is missing or stale. Calibration was treated as a core requirement for decision integration, motivating reliability corrections such as temperature scaling and prior-sensitive adjustments, and clarifying how numerical miscalibration translates into unstable thresholds and unpredictable alert volumes. We further discussed uncertainty quantification and selective prediction as control primitives that can reduce system-level risk under out-of-distribution conditions, and we outlined evaluation and monitoring protocols that separate genuine information gains from base-rate artifacts and that operate under delayed labels. Finally, we addressed systems constraints for streaming inference, including causal data access, incremental computation, versioned calibration, and reproducible audit logging. Taken together, these components provide a principled pathway from multimodal encoders to operationally meaningful risk scores, with reliability expressed as a property that is designed, measured, and maintained rather than assumed [32].

## References


[1] S. Qiao, G. Khushf, X. Li, J. Zhang, and B. Olatosi, "Developing an ethical framework-guided instrument for assessing bias in ehr-based big data studies: a research protocol.," BMJ open, vol. 13, pp. e070870–e070870, 8 2023.

[2] W. P. Newton, B. Thompson, T. L. Campbell, and D. C. Spencer, "Keeping our eye on the ball: managing the evolution of electronic health records.," Annals of family medicine, vol. 4, pp. 184–185, 3 2006.

[3] L. J. Liu, V. Ortiz-Soriano, J. A. Neyra, and J. Chen, "Kit-lstm: Knowledge-guided time-aware lstm for continuous clinical risk prediction," 11 2022.

[4] S. Hoffman, Electronic Health Records and Medical Big Data. Cambridge University Press, 12 2016.

[5] A. C. Cheng, M. K. Banasiewicz, J. D. Johnson, L. Sulieman, N. Kennedy, F. Delacqua, A. A. Lewis, M. M. Joly, A. J. Bistran-Hall, S. Collins, W. H. Self, M. S. Shotwell, C. J. Lindsell, and P. A. Harris, "Evaluating automated electronic case report form data entry from electronic health records.," Journal of clinical and translational science, vol. 7, pp. e29–e29, 12 2022.

[6] S. S. Sebok-Syer, A. M. Dukelow, R. Sedran, L. Shepherd, A. McConnell, J. M. Shaw, and L. Lingard, "Developing and authenticating an electronic health record-based report card for assessing residents' clinical performance.," AEM education and training, vol. 7, pp. e10851–e10851, 3 2023.

[7] P. S. Sen and N. Mukherjee, "A case study on implementation of health data standards for smart healthcare," International Journal of Engineering Research in Computer Science and Engineering, vol. 9, pp. 17–25, 6 2022.

[8] H. Marani, I. Halperin, T. Jamieson, and G. Mukerji, "Quality gaps of electronic health records in diabetes care.," Canadian journal of diabetes, vol. 44, pp. 350–355, 11 2019.

[9] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications.," NPJ digital medicine, vol. 6, pp. 98–98, 5 2023.

[10] J. Qu, "Security research of blockchain technology in electronic medical records.," Medicine, vol. 101, pp. e30507–e30507, 9 2022.

[11] "A standard model for evaluating return on investment from electronic health record implementation," 1 2014.

[12] S. R. Maloney, S. Peterson, A. M. Kao, W. C. Sherrill, J. M. Green, and G. Sachdev, "Surgery resident time consumed by the electronic health record.," Journal of surgical education, vol. 77, pp. 1056–1062, 4 2020.

[13] R. A. Ridberg, A. L. Yaroch, N. B. Nugent, C. B. Shanks, and H. Seligman, "A case for using electronic health record data in the evaluation of produce prescription programs.," Journal of primary care & community health, vol. 13, pp. 21501319221101849–21501319221101849, 5 2022.

[14] B. Sadanandan, "Multimodal deep learning for early prediction of patient deterioration in the icu: Integrating time-series ehr data with clinical notes," Journal of Data Science, Predictive Analytics, and Big Data Applications, vol. 10, no. 7, pp. 1–21, 2025.

[15] P. Otero, P. Durán, D. Setton, A. Eymann, J. Busaniche, and J. Llera, "Mismatch between the prevalence of overweight and obese children and adolescents and recording in electronic health records: a cross-sectional study.," Informatics in primary care, vol. 19, pp. 75–82, 5 2011.

[16] C. L. M. Joseph, A. Tang, D. W. Chesla, M. M. Epstein, P. A. Pawloski, A. B. Stevens, S. C. Waring, B. K. Ahmedani, C. C. Johnson, and C. D. Peltz-Rauchman, "Demographic differences in willingness to share electronic health records in the all of us research program.," Journal of the American Medical Informatics Association : JAMIA, vol. 29, pp. 1271–1278, 4 2022.

[17] R. W. Hilliard, J. Haskell, and R. Gardner, "Are specific elements of electronic health record use associated with clinician burnout more than others," Journal of the American Medical Informatics Association : JAMIA, vol. 27, pp. 1401–1410, 7 2020.

[18] R. Shaw, M. Stroo, C. Fiander, and K. McMillan, "Selecting mobile health technologies for electronic health record integration: Case study (preprint)," 8 2020.

[19] G. Strudwick, T. Tajirian, J. Kemp, N. Coombe, U. Haider, S. Kaur, S. Murphy, H. D. Shin, S. Ling, and D. Jankowicz, "Utilizing an informatics engagement strategy as an approach to sustain and retain the nursing workforce.," Nursing leadership (Toronto, Ont.), vol. 35, pp. 42–54, 1 2023.

[20] S. Müller, "Is there a civic duty to support medical ai development by sharing electronic health records?," BMC medical ethics, vol. 23, pp. 134–134, 12 2022.

[21] K. L. Walker, O. Kirillova, S. Gillespie, D. Hsiao, V. Pishchalenko, A. K. Pai, J. Puro, R. Plumley, R. Kudyakov, W. Hu, A. Allisany, M. A. McBurnie, S. E. Kurtz, and B. Hazlehurst, "Using the cer hub to ensure data quality in a multi-institution smoking cessation study," Journal of the American Medical Informatics Association : JAMIA, vol. 21, pp. 1129–1135, 7 2014.

[22] D. Ko, N. Forrest, Q. Mai, A. Pawlowski, K. Balsley, A. Chung, R. Ramsey-Goldman, and T. L. Walunas, "Electronic health record data use in the assessment of quality indicators for glucocorticoid osteoporosis screening in systemic lupus erythematosus.," Lupus, vol. 31, pp. 1516–1522, 7 2022.

[23] A. Zalin, C. Papoutsi, K. Shotliff, A. Majeed, C. Marston, and J. E. Reed, "The use of information for diabetes research and care: Patient views in west london," Practical Diabetes, vol. 33, pp. 81–, 4 2016.

[24] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, "Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data," Journal of the American Medical Informatics Association : JAMIA, vol. 27, pp. 1921–1934, 10 2020.

[25] V. Palahin, O. Palahina, and V. Hahen, "Determination of efficiency criteria in the development of medical information systems," HERALD OF KHMELNYTSKYI NATIONAL UNIVERSITY, vol. 297, pp. 116–123, 7 2021.

[26] Y. Liu, W. Wang, G. G. Gao, and R. Agarwal, "Echoes of biases: How stigmatizing language affects ai performance," 5 2023.

[27] M. J. Francois and E. E. Obisike, "Accelerating the national implementation of electronic health records in canada," European Scientific Journal, ESJ, vol. 12, pp. 65–65, 5 2016.

[28] R. B. Parikh, J. Schnall, M. Liu, P. Gabriel, C. Chivers, M. Draugelis, W. Ferrell, C. Lerman, J. E. Bekelman, and J. Chen, "Augmenting machine learning algorithms to predict mortality using patient-reported outcomes in oncology.," Journal of Clinical Oncology, vol. 39, pp. 1510–1510, 5 2021.

[29] N. C. Apathy, L. Rotenstein, D. W. Bates, and A. J. Holmgren, "Documentation dynamics: Note composition, burden, and physician efficiency.," Health services research, vol. 58, pp. 674–685, 11 2022.

[30] K. Adane, M. Gizachew, and S. Kendie, "The role of medical data in efficient patient care delivery: a review.," Risk management and healthcare policy, vol. 12, pp. 67–73, 4 2019.

[31] A. Sinha, L. A. Stevens, F. Su, N. M. Pageler, and D. S. Tawfik, "Measuring electronic health record use in the pediatric icu using audit-logs and screen recordings.," Applied clinical informatics, vol. 12, pp. 737–744, 8 2021.

[32] R. A. Hubbard, J. Chubak, and C. M. Rutter, "Estimating screening test utilization using electronic health records data.," EGEMS (Washington, DC), vol. 2, pp. 14–14, 11 2014.