PRISM: Calibrated Bayesian Fusion and Auditable Attribution for Reliable LLM Event Prediction from Text

Chengyuan Jin[1,2] **Daojian Zeng**[4] **Kang Liu**[1,2] **Jun Zhao**[1,2] **Yubo Chen**[3,*]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[2]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]School of Information Engineering, Minzu University of China
[4]Hunan Normal University, China
[*]yubo.chen@muc.edu.cn

Abstract

We study event prediction from news text when intermediate drivers are not expert-annotated. Directly prompting large language models (LLMs) for forecast probabilities can yield miscalibrated outputs and explanations that are hard to audit. We propose the Probabilistic Reliability and Interpretability System, PRISM, which separates semantic extraction from probabilistic decision making. PRISM uses LLMs only to extract interpretable factors, treats repeated extractions as noisy measurements of latent states, and performs Bayesian inference to produce posterior predictive probabilities that propagate extraction and measurement uncertainty. For empirical reliability under temporal dependence, we add low-capacity post-hoc calibration and time-adaptive conformal prediction and report calibration and coverage diagnostics. PRISM also reports factor-contrast association summaries with uncertainty and explicit interpretation boundaries. On a UCDP–GDELT conflict benchmark, PRISM achieves AUROC 0.821 and ECE 0.058 on the test set; at 90% target coverage, it attains empirical coverage 0.908 with average set size 1.38.

1 INTRODUCTION

Event prediction from news text is central to high-stakes decision making. Stakeholders need accurate forecasts, trustworthy uncertainty, and a clear link from predictions back to textual evidence. Recent work shows that large language models (LLMs) can be competitive forecasters when directly elicited for probabilities or used in tool-augmented pipelines [Schoenegger and Park, 2023, Halawi et al., 2024, Hsieh et al., 2024, Guan et al., 2024]. Yet many important settings remain weakly supervised: outcomes may come from event logs, intermediate drivers are not expert-annotated, and the only timely evidence is unstructured text.

Existing approaches largely follow two patterns. One directly asks an LLM for a probability and a rationale [Schoenegger and Park, 2023, Halawi et al., 2024], optionally aggregating prompts or seeds to reduce variance [Qiang et al., 2024, Tonolini et al., 2024]. Another extracts structured signals from text and trains a predictive model, but extraction is often treated as ground truth, with limited treatment of measurement error and limited auditability of how evidence maps to forecasts [Ludwig et al., 2024, Zhang and Martinez, 2025, Nafar et al., 2024].

A first obstacle is *probabilistic reliability*. An LLM's printed number depends on token-level likelihoods and prompt phrasing, and it can be miscalibrated and unstable even under stationary base rates [Nafar et al., 2024, Guo et al., 2017]. Temporal dependence and nonstationarity further weaken standard uncertainty assumptions [Barber et al., 2022, Xu and Xie, 2021]. Without an explicit probabilistic semantics that separates latent world states from text-conditioned extraction noise, it is unclear what the reported "probability" means or how to propagate uncertainty from extraction to prediction.

A second obstacle is *auditability and interpretation boundaries*. Free-form rationales are not reproducible quantitative objects: they rarely specify which text-derived drivers were used, how strongly they mattered, or how driver uncertainty affects the forecast. Turning text-derived factors into intervention claims also requires identification assumptions that should be stated and stress-tested, not implied by a generated explanation [Imbens and Rubin, 2015, Robins, 1986]. In high-stakes settings, we therefore need controlled predictive associations with diagnostics and clear interpretation boundaries.

We address these challenges with the **Probabilistic Reliability and Interpretability System (PRISM)**, which separates semantic extraction from probabilistic decision making. PRISM uses the LLM as a *factor sensor*: it extracts

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2595–2617.

a fixed inventory of interpretable factors and evidence spans, and we repeat extraction to obtain noisy replicate measurements. We fuse replicates with a rater-style hierarchical measurement model [Dawid and Skene, 1979, Hui and Walter, 1980] and couple latent factors to outcomes in a joint Bayesian model, yielding posterior predictive probabilities that propagate extraction uncertainty end-to-end. For empirical reliability under temporal dependence, PRISM adds low-capacity post-hoc calibration [Platt, 1999, Zadrozny and Elkan, 2002, Guo et al., 2017] and time-adaptive conformal prediction [Vovk et al., 2005, Barber et al., 2022, Xu and Xie, 2021]. PRISM also produces an auditable quantitative report via factor-contrast summaries with uncertainty and explicit interpretation boundaries.

We evaluate PRISM on a conflict forecasting benchmark built from UCDP–GED outcomes and GDELT news text [Sundberg and Melander, 2013, Leetaru and Schrodt, 2013], demonstrating improved calibration and uncertainty control over direct LLM forecasting while providing transparent, traceable factor-based reports. Our contributions are:

- **Measurement-to-forecast semantics:** We treat LLM extraction as noisy repeated measurements of interpretable latent factors with evidence spans, and perform joint Bayesian inference so extraction uncertainty propagates to posterior predictive probabilities and quantitative summaries.
- **Reliability under temporal dependence:** We complement posterior predictive inference with low-capacity calibration and time-adaptive conformal prediction, and report calibration and coverage diagnostics.
- **Auditable quantitative reporting:** We provide factor-contrast association summaries with uncertainty and explicit interpretation boundaries, together with diagnostics that discourage causal over-claims.

2 RELATED WORK

LLM event prediction and Bayesian fusion. Forecasting benchmarks and tournaments highlight both the promise of LLMs for event prediction and their probability-level brittleness and miscalibration under direct elicitation [Schoenegger and Park, 2023, Schoenegger et al., 2024, Halawi et al., 2024, Karger et al., 2024, Guan et al., 2024]. ForecastBench, for example, evaluates forecasters on unresolved future questions to reduce leakage risk, whereas our setting assumes temporally aligned document sets for each forecasting unit. Beyond text, tabular foundation models show that strong predictions can be achieved in low-data regimes by pretraining on diverse tables [Hollmann et al., 2025]; PRISM is complementary in that it targets text-conditioned latent drivers without intermediate labels and emphasizes auditability of the full measurement-to-forecast pipeline. Prompt ensembles and related decision-theoretic views aim to reduce variance or improve decisions by aggregating multiple LLM calls [Tonolini et al., 2024, Vashurin et al., 2025]; PRISM differs by restricting the LLM to extracting interpretable factors with evidence and by pushing probabilistic inference into an explicit model whose uncertainty can be audited. Classical repeated-rater, truth-inference, and Bayesian classifier-combination models, including independent Bayesian classifier combination (IBCC), fuse noisy measurements via explicit confusion or reliability parameters [Dawid and Skene, 1979, Hui and Walter, 1980, Simpson et al., 2012, Li et al., 2019]; PRISM adapts this perspective to same-template LLM prompt replicates, uses a low-dimensional prompt-dependence correction rather than persistent worker heterogeneity or worker-community structure, and couples measurement to forecasting through a single joint model.

Calibration and risk control. Post-hoc calibration, such as Platt scaling and isotonic regression, improves alignment to empirical frequencies under limited capacity [Platt, 1999, Zadrozny and Elkan, 2002, Guo et al., 2017, Kuleshov et al., 2018]. Conformal prediction provides uncertainty sets under exchangeability [Vovk et al., 2005] and has extensions targeting dependence or nonstationarity [Barber et al., 2022, Xu and Xie, 2021]. PRISM treats these as complementary layers: posterior predictive inference provides coherent probabilities under an explicit model, while calibration and rolling conformal procedures provide empirical reliability objectives under temporal dependence.

Attribution and causality from text. Intervention-style interpretation requires explicit identification assumptions and careful diagnostics and sensitivity analyses to clarify what can and cannot be concluded from observational text-derived factors [Robins, 1986]. For causal discovery, classical constraint-based methods such as PC and FCI [Spirtes et al., 2000, 2013] motivate recent efforts that incorporate retrieval-augmented LLM workflows [Zhang et al., 2024], while other work highlights practical limitations and calls for restricted use [Wu et al., 2025]. PRISM focuses on auditable factor extraction and uncertainty-aware predictive attribution rather than full causal discovery.

3 PROBLEM SETUP

We study binary event prediction from text at country–time resolution with latent intermediate factors. Let c index countries and $t \in \{1, \dots, T\}$ time bins (e.g., day or month). At each (c, t) we observe documents $D_{c,t} = \{d_{c,t,1}, \dots, d_{c,t,n_{c,t}}\}$ and predict a binary event indicator $Y_{c,t+\Delta} \in \{0, 1\}$ at horizon Δ. Outcomes come from event logs or automatic matching, but intermediate factor states are not expert-annotated. We suppress c when unambiguous.

Goal. For each t, output $\hat{p}_t \approx P(Y_{t+\Delta} = 1 \mid D_{\leq t})$ with

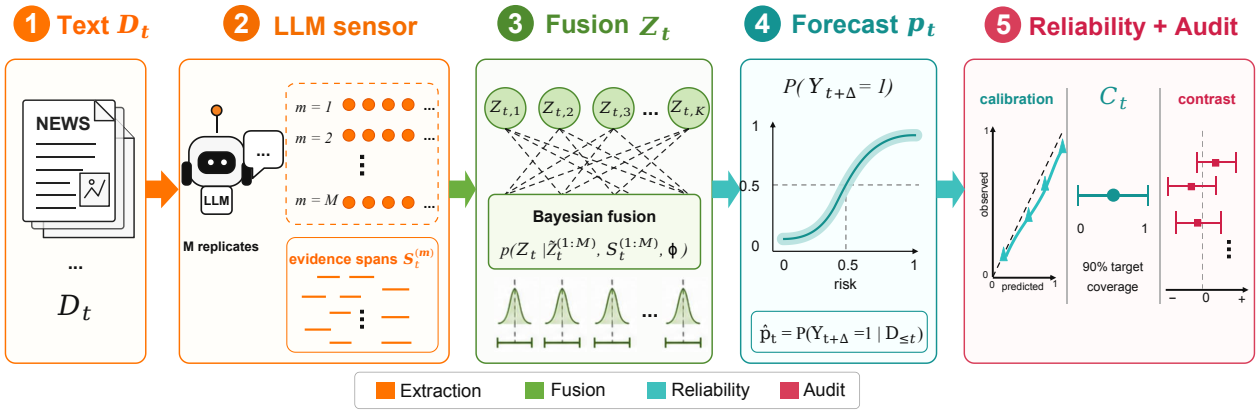

Figure 1: End-to-end PRISM pipeline. Documents D_t are converted into repeated, evidence-grounded factor measurements $\tilde{Z}_t^{(m)}$ and spans $S_t^{(m)}$; the fusion layer infers latent factors Z_t under replicate noise and dependence; the outcome layer computes posterior predictive risk $\hat{p}_t$; and the final layer applies calibration, conformal prediction, and factor-contrast diagnostics for reliability and auditability.

coherent probabilistic semantics and empirical reliability, together with an auditable report that links forecasts to text-derived factors with uncertainty.

3.1 ASSUMPTIONS AND SCOPE

PRISM is designed to make modeling assumptions explicit and auditable, rather than implicit in prompting. Main assumptions: **Repeated measurement adequacy:** prompt measurements $\tilde{Z}_{t,k}^{(m)}$ follow a rater-style model under an orientation convention that keeps factor meanings stable. In real-data experiments we use no anchor labels, rely on a weak orientation prior, and treat anchors as an optional sensitivity check in semi-synthetic validation. **Time-adaptive uncertainty control:** rolling and sequential conformal prediction is evaluated as an empirical coverage objective under temporal dependence, not a finite-sample distribution-free guarantee. **Interpretation boundaries:** quantitative factor contrasts are controlled predictive associations by default; an intervention-style reading requires additional identification assumptions (e.g., sequential unconfoundedness and positivity) and should be interpreted with diagnostics and sensitivity analyses.

4 PRISM

4.1 OVERVIEW

Running Example. Consider predicting *civil unrest* $Y_{t+\Delta}$ in a country. From daily news D_t, we extract latent factors Z_t such as *political protests*, *economic sanctions*, and *government crackdowns*. Since factor states are not labeled, we query an LLM M times to obtain noisy measurements $\tilde{Z}_t^{(m)}$, fuse them into a posterior over Z_t, and learn their association with $Y_{t+\Delta}$.

PRISM has three stages: (i) document-conditioned extraction of a fixed factor inventory with evidence spans, (ii) Bayesian fusion and posterior predictive forecasting under an explicit generative model, and (iii) empirical reliability control via calibration and time-adaptive conformal prediction. A reporting stage outputs factor-contrast summaries with uncertainty and explicit interpretation boundaries.

We define the joint generative model underlying PRISM, which unifies semantic extraction, forecasting, and quantitative reporting:

$$\begin{aligned} & p(Y_{1:T-\Delta}, Z_{1:T}, \tilde{Z}_{1:T}, \theta, \phi) \\ &= \prod_{t=1}^{T-\Delta} p(Y_{t+\Delta} \mid Z_t, E_t, \theta) \\ &\quad \cdot \prod_{t=1}^{T} \left[p(Z_t \mid \phi) \prod_{m=1}^{M} p(\tilde{Z}_t^{(m)} \mid Z_t, \phi) \right] \\ &\quad \cdot p(\theta)\, p(\phi). \end{aligned} \tag{1}$$

where $\phi = (\pi, \rho, \kappa)$ are measurement parameters and $\theta = (\alpha, \beta, w)$ outcome parameters. We fix concentration τ_z for the Beta prior on Z; E_t are optional text-derived covariates; and $\rho = \{(\rho_k^{(1)}, \rho_k^{(0)})\}_{k=1}^{K}$ collects factor-specific sensitivity and specificity. This single joint model supports extraction with uncertainty, posterior predictive forecasting, calibration/conformal reliability layers, and factor-contrast reporting with interpretation boundaries. In our experiments, E_t is the mean-pooled RoBERTa document embedding of D_t (Appendix D.2).

Core priors and definitions are summarized in Appendix E.2.

For scalability, we learn an approximate posterior over global measurement parameters on a stratified replicate subset and propagate it downstream via Monte Carlo; details, pseudocode, and audits are in Appendix B, Appendix D.1, and Appendix C.1. For $t \in \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}}$, the posterior predictive expectation integrates over posterior draws of θ and $p(Z_t \mid \tilde{Z}_t, \phi^{(l)})$ under each draw $\phi^{(l)} \sim q(\phi)$.

4.2 LLM AS A NOISY FACTOR SENSOR

We fix K interpretable factors and represent their latent *intensities* by $Z_t = (Z_{t,1}, \ldots, Z_{t,K})$, where each $Z_{t,k} \in (0, 1)$ encodes factor presence strength at time t.

Design and efficiency. We obtain M conditionally exchangeable repeated measurements by extracting all K factors in a single structured LLM call per replicate (fixed template; different random seeds), and we use a partial replicate design that repeats extraction only on a stratified subset to estimate global measurement parameters and propagate them downstream; details are in Appendix D.1 and Appendix D.1.

PRISM queries an LLM to obtain noisy measurements $\tilde{Z}_t^{(m)}$ and evidence spans $S_t^{(m)}$ grounded in D_t. We treat $\tilde{Z}_t^{(m)} = (\tilde{Z}_{t,1}^{(m)}, \ldots, \tilde{Z}_{t,K}^{(m)})$ as repeated noisy observations of Z_t and fuse them with a hierarchical measurement model. For binary factors, a baseline instance is

$$Z_{t,k} \mid \pi_k, \tau_z \sim \text{Beta}\big(\tau_z \pi_k, \tau_z(1-\pi_k)\big), \tag{2}$$

$$\tilde{Z}_{t,k}^{(m)} \mid Z_{t,k} \sim \text{Bernoulli}(\eta_{t,k}), \tag{3}$$

$$\eta_{t,k} = \rho_k^{(1)} Z_{t,k} + (1 - \rho_k^{(0)})(1 - Z_{t,k}), \tag{4}$$

where $\rho_k^{(1)} = P(\tilde{Z} = 1 \mid Z = 1)$ is sensitivity and $\rho_k^{(0)} = P(\tilde{Z} = 0 \mid Z = 0)$ specificity. This separates the latent world state $Z_{t,k}$ from the LLM report $\tilde{Z}_{t,k}^{(m)}$ for uncertainty propagation; the symmetric flip-noise model is the special case $\rho_k^{(1)} = \rho_k^{(0)}$.

Conditional dependence across prompts. Classical repeated-measurement models assume conditional independence given $Z_{t,k}$, but LLM prompts can share document-conditioned biases. We therefore include an exchangeable correlated-measurement extension with an explicit cross-prompt dependence parameter and validate it with agreement-based posterior predictive checks.

Exchangeability diagnostics across prompt replicates. Exchangeability can fail if replicate index (e.g., seed) induces systematic differences even under fixed decoding. We report replicate-index diagnostics on the replicate subset; when they fail for a factor, we inflate its measurement uncertainty, avoid strong semantic claims in attribution, and rely more on calibration/conformal layers for decision uncertainty.

Identifiability, orientation, and anchors. The measurement layer has label-switching and orientation ambiguity, so we use a weak orientation convention with optional anchors in semi-synthetic checks and treat identifiability as an auditable condition.

Beyond binary factors. Many factors are ordinal or continuous. PRISM supports non-binary factors by swapping measurement and outcome links (e.g., ordered logistic for ordinal Z, Gaussian for continuous Z) while preserving the same posterior predictive semantics.

4.3 BAYESIAN POSTERIOR PREDICTIVE FORECASTING

Given latent factors Z_t and optional text-derived covariates E_t (e.g., embeddings), we model the event probability using a Bayesian generalized linear model:

$$P(Y_{t+\Delta} = 1 \mid Z_t, E_t, \theta) = \sigma(\alpha + \beta^\top Z_t + w^\top E_t), \tag{5}$$

where $\sigma(\cdot)$ is the logistic sigmoid. The key modeling choice is that measurement and outcome layers form a single joint generative model. Let $\tilde{Z}_{1:T} = \{\tilde{Z}_t^{(m)}\}_{t,m}$ denote all prompt measurements and $\phi = (\pi, \rho, \kappa)$ measurement parameters. The joint posterior is

$$\begin{aligned} & p(Z_{1:T}, \phi, \theta \mid \tilde{Z}_{1:T}, E_{1:T}, Y_{1:T-\Delta}) \\ \propto\ & p(Y_{1:T-\Delta} \mid Z_{1:T}, E_{1:T}, \theta) \\ & \cdot p(\tilde{Z}_{1:T} \mid Z_{1:T}, \phi) \cdot p(Z_{1:T} \mid \phi) \\ & \cdot p(\phi) \cdot p(\theta). \end{aligned} \tag{6}$$

which propagates extraction and measurement uncertainty into forecasts and contrast summaries. We use weakly regularizing priors on $\theta = (\alpha, \beta, w)$ to avoid overconfident posteriors under rare events and limited supervision. The predictive probability integrates uncertainty from both extraction and parameters:

$$\begin{aligned} \hat{p}_t &= P(Y_{t+\Delta} = 1 \mid D_{\le t}) \\ &= \int P(Y_{t+\Delta} = 1 \mid Z_t, E_t, \theta)\, \Pi_t(dZ_t, d\theta). \end{aligned} \tag{7}$$

Here Π_t denotes the posterior measure induced by $p(Z_t, \theta \mid \tilde{Z}_{1:t}, E_{1:t}, Y_{1:t-\Delta})$. This yields posterior predictive distributions rather than point estimates, supporting auditing, uncertainty reporting, and sensitivity analysis.

Avoiding text-factor collinearity. Because Z_t and E_t are both derived from the same documents, we treat E_t as optional and either set $w = 0$ (factors-only) or use split-respecting residualized embeddings.

Prior choice and sensitivity. We use weakly regularizing priors and report sensitivity analyses over key prior settings.

Table 1: Forecast accuracy and calibration on the UCDP–GDELT benchmark. Values are mean ± std over country-level bootstrap resamples on the test set. Post-hoc isotonic calibration is monotone and does not change ranking, so AUROC/AUPRC are unchanged for calibrated variants. Statistical uncertainty for method differences is assessed via country-level paired bootstrap intervals; see Appendix D.4. * marks that PRISM improves over Direct-LLM (Llama-3-70B) at $p < 0.05$ under a conservative normal approximation using paired country-level bootstrap resamples.

Method	NLL↓	Brier↓	ECE↓	AUROC↑	AUPRC↑
History+Controls Logistic (w/ lagged Y)	0.263 ± 0.019	0.084 ± 0.011	0.091 ± 0.013	0.761 ± 0.016	0.290 ± 0.018
Supervised Text Model (DeBERTa-v3)	0.246 ± 0.015	0.075 ± 0.006	0.098 ± 0.009	0.773 ± 0.014	0.310 ± 0.021
Direct-LLM (Llama-3-70B)	0.288 ± 0.023	0.094 ± 0.009	0.176 ± 0.017	0.742 ± 0.018	0.220 ± 0.024
Direct-LLM + Isotonic Calib.	0.241 ± 0.014	0.070 ± 0.007	0.072 ± 0.008	0.742 ± 0.018	0.220 ± 0.024
LLM-Ensemble ($M = 5$)	0.249 ± 0.017	0.076 ± 0.008	0.118 ± 0.011	0.781 ± 0.013	0.260 ± 0.016
Text-Only Predictor (RoBERTa)	0.256 ± 0.019	0.081 ± 0.007	0.105 ± 0.010	0.724 ± 0.019	0.240 ± 0.023
Two-Stage (Dawid–Skene + Bayes)	0.234 ± 0.013	0.067 ± 0.005	0.085 ± 0.009	0.803 ± 0.012	0.325 ± 0.015
PRISM (Ours)*	$\mathbf{0.212} \pm 0.012$	$\mathbf{0.062} \pm 0.004$	$\mathbf{0.058} \pm 0.006$	$\mathbf{0.821} \pm 0.018$	$\mathbf{0.345} \pm 0.020$

Posterior inference. We use scalable approximate inference for the main runs and audit uncertainty summaries against an HMC/NUTS verification subset.

4.4 CALIBRATION AND TIME-ADAPTIVE CONFORMAL PREDICTION

The posterior predictive probability $\hat{p}_t = P(Y_{t+\Delta} = 1 \mid D_{\le t})$ is coherent under the assumed model, but misspecification, distribution shift, and imperfect extraction can cause miscalibration. PRISM separates *probabilistic coherence* from *empirical reliability*: we optionally fit a low-capacity monotone calibration map $c(\cdot)$ on a held-out calibration split and set $\hat{p}_t^{\text{cal}} = c(\hat{p}_t)$. This is an explicit model-mismatch correction (not a guarantee of posterior semantics preservation). We report calibrated and uncalibrated results, evaluating calibration via ECE and reliability diagrams.

To control uncertainty under dependence, we apply time-adaptive conformal prediction on top of calibrated scores using rolling or sequential windows.

4.4.1 Time-Adaptive Conformal Prediction

Standard split conformal requires exchangeability, which fails in time series. We therefore use a **rolling window or sequential conformal** protocol and report empirical coverage over time. For a candidate label $y \in \{0, 1\}$, define the nonconformity score

$$s_t(y) = 1 - \hat{p}_t^{\text{cal}}(y), \tag{8}$$

where $\hat{p}_t^{\text{cal}}(1) = \hat{p}_t^{\text{cal}}$ and $\hat{p}_t^{\text{cal}}(0) = 1 - \hat{p}_t^{\text{cal}}$. On the calibration split, we compute scores $s_i(Y_i)$ and set the threshold $q_{1-\alpha}$ as the $(1-\alpha)$ quantile. The prediction set is

$$C_t = \{y \in \{0, 1\} : s_t(y) \le q_{1-\alpha}\}. \tag{9}$$

We use a rolling implementation with class-conditional thresholds and gap windows, evaluating it as an empirical coverage objective under dependence.

4.5 AUDITABLE QUANTITATIVE ATTRIBUTION UNDER A LATENT MODEL

PRISM provides auditable *quantitative attribution reports* with explicit interpretation boundaries. Treating factor k as $T_t = Z_{t,k}$ and the remaining variables as $X_t = (Z_{t,-k}, E_t)$, our primary summary is an in-support quantile contrast,

$$\begin{aligned}\tau_k^{\text{Q}} = \mathbb{E}\Big[&\mathbb{E}[Y \mid T = q_{0.9}(T), X] \\ &- \mathbb{E}[Y \mid T = q_{0.1}(T), X]\Big],\end{aligned} \tag{10}$$

where $q_{0.9}(T)$ and $q_{0.1}(T)$ are posterior support quantiles for the latent factor intensity. This choice avoids making the main attribution summary depend on unsupported boundary values when posterior mass is concentrated away from 0 or 1. We also report the standardized extreme contrast $\tau_k = \mathbb{E}[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]]$ as a secondary absent-versus-fully-present summary under the fitted outcome model. We interpret both contrasts as *controlled predictive associations* by default, not causal effects, unless explicit identification assumptions (e.g., consistency, positivity/overlap, no unmeasured confounding given X, no interference, adequate measurement of T) and diagnostics support such a reading. In our reporting, we emphasize overlap checks, placebo tests, and sensitivity analysis, and we make all interpretation boundaries explicit.

4.6 AUDITABLE OUTPUT

For each forecast, PRISM outputs an audit bundle: extracted factor measurements with evidence spans, posterior predictive distributions, calibration maps, conformal thresholds, and attribution reports with uncertainty. Here "auditable" means each number is traceable to an explicit model component and a well-defined posterior predictive (or posterior-averaged) computation, and can be inspected alongside the text evidence; it does not imply a causal claim. This supports reproducibility, manual review, and post-hoc error analysis.

5 EXPERIMENTS

5.1 EXPERIMENTAL QUESTIONS

We evaluate PRISM along four questions: whether it improves forecast quality over Direct-LLM and supervised text predictors, whether it improves empirical reliability through calibration and conformal coverage, whether it reduces sensitivity to prompt and phrasing perturbations, and whether it supports stable quantitative attribution summaries. We validate attribution on semi-synthetic experiments and stress tests, and audit it on real data. Complete experimental protocol and implementation details are provided in Appendix B and Appendix D.1.

5.2 BENCHMARK

We evaluate PRISM on UCDP–GDELT, combining English news from GDELT [Leetaru and Schrodt, 2013] with conflict outcomes from UCDP GED [Sundberg and Melander, 2013]. We aggregate documents to country-month bins $D_{c,t}$ and forecast next-month outcomes with horizon $\Delta = 1$: $Y_{c,t+1} = 1$ if at least one qualifying state-based violence event occurs in country c during month $t + 1$. We use fixed train/calibration/test splits. Preprocessing, filtering, and control alignment are in Appendix D.1. Our main specification uses a global measurement layer and country-specific outcome parameters; the partial replicate design and inference details are in Appendix D.1.

Factor definition. We fix a compact factor inventory ($K = 8$) with operational definitions before evaluation. The locked inventory and definitions used throughout all methods and audits are in Appendix B.1 (Table 7).

Factor semantics spot-check. As a spot-check of factor semantics on real documents, we sample 100 document–factor pairs and compare PRISM's posterior-mode label against a human binary annotation using the factor definitions. The agreement rate is 83% (83/100) (95% CI: $[75\%, 90\%]$), with Cohen's $\kappa = 0.68$; see Appendix A.6.

Benchmark scope. We report results on one demanding benchmark, UCDP–GDELT, to keep the factor inventory, prompt protocol, temporal splits, and audit procedure fixed while evaluating calibration under rare events and temporal dependence. This experimental section should therefore be read as a controlled conflict-forecasting study rather than as a cross-domain benchmark; Section 6 discusses the broader transfer limitations.

5.3 BASELINES

We compare PRISM against Direct-LLM forecasting (with/without post-hoc calibration), prompt-matched LLM self-ensembles, a two-stage measurement-prediction pipeline, supervised text models, history+controls baselines, and conduct PRISM component ablations. Baseline and protocol details are in Appendix B.2.

5.4 METRICS

We report forecast quality (NLL, Brier, AUROC, AUPRC), calibration (ECE and reliability diagrams), and conformal reliability (coverage and set size). For attribution, we report contrast error and interval coverage on semi-synthetic ground truth and diagnostics on real data. Metric definitions and aggregation conventions are in Appendix C.5 and Appendix C.4.

5.5 IMPLEMENTATION SUMMARY

We use a compact factor budget and a fixed replicate budget, with defaults $K = 8$ and $M = 5$. All LLM-based methods use matched backbone/decoding settings to isolate the effect of PRISM's statistical layer, and all tuning is performed on the calibration split only (test labels are used only once under a frozen configuration). Full preprocessing, replicate design, and prompt details are in Appendix D.1.

Table 2: Ablations for M and K. Results show diminishing returns beyond $M = 5$ and $K = 8$.

Setting	NLL↓	Brier↓	ECE↓	AUROC↑	AUPRC↑
Replicates M					
$M = 1$	0.255 ± 0.017	0.080 ± 0.008	0.118 ± 0.011	0.788 ± 0.019	0.292 ± 0.023
$M = 3$	0.222 ± 0.014	0.065 ± 0.005	0.065 ± 0.007	0.817 ± 0.017	0.335 ± 0.020
$M = 5$ (Main)	0.212 ± 0.012	0.062 ± 0.004	0.058 ± 0.006	0.821 ± 0.018	0.345 ± 0.020
$M = 7$	0.211 ± 0.012	0.060 ± 0.004	0.056 ± 0.005	0.823 ± 0.017	0.347 ± 0.019
Factors K					
$K = 4$	0.238 ± 0.016	0.072 ± 0.007	0.075 ± 0.009	0.780 ± 0.021	0.285 ± 0.023
$K = 6$	0.219 ± 0.013	0.065 ± 0.005	0.062 ± 0.007	0.810 ± 0.018	0.328 ± 0.020
$K = 8$ (Main)	0.212 ± 0.012	0.062 ± 0.004	0.058 ± 0.006	0.821 ± 0.018	0.345 ± 0.020
$K = 10$	0.211 ± 0.012	0.062 ± 0.004	0.057 ± 0.007	0.822 ± 0.017	0.346 ± 0.021
$K = 12$	0.213 ± 0.013	0.063 ± 0.005	0.059 ± 0.006	0.819 ± 0.019	0.342 ± 0.021

5.6 HYPERPARAMETER ABLATIONS

We include a main text ablation on the two core PRISM hyperparameters: the number of repeated measurements M, which controls extraction uncertainty resolution and measurement layer identifiability, and the number of factors K, which controls representation capacity and attribution granularity. The protocol holds the split, decoding settings, and calibration and conformal procedures fixed, and varies only the target hyperparameter. Table 2 reports the results.

Table 3: LLM backbone sensitivity. PRISM maintains a performance advantage across model scales, though stronger backbones benefit all methods.

Backbone	Method	AUROC↑	AUPRC↑	ECE↓
Llama-3-8B	Direct-LLM	0.680 ± 0.022	0.185 ± 0.024	0.250 ± 0.015
	LLM-Ensemble	0.705 ± 0.019	0.205 ± 0.021	0.210 ± 0.014
	PRISM	0.790 ± 0.016	0.295 ± 0.019	0.080 ± 0.009
Llama-3-70B	Direct-LLM	0.742 ± 0.018	0.220 ± 0.022	0.176 ± 0.012
	LLM-Ensemble	0.781 ± 0.015	0.260 ± 0.018	0.118 ± 0.010
	PRISM	0.821 ± 0.018	0.345 ± 0.020	0.058 ± 0.006
GPT-4o	Direct-LLM	0.780 ± 0.015	0.255 ± 0.020	0.120 ± 0.010
	LLM-Ensemble	0.805 ± 0.014	0.290 ± 0.018	0.095 ± 0.008
	PRISM	0.835 ± 0.015	0.365 ± 0.018	0.052 ± 0.006

Table 4: Ablations of PRISM on reliability components. Values are mean ± std over country-level bootstrap resamples on the test set, except for Coverage and Set Size which are reported as pooled aggregates. Target coverage is 90%. Dashes indicate metrics not defined when conformal prediction is disabled.

Variant	NLL↓	ECE↓	Cov.↑	Set Size↓
w/o measurement	0.231 ± 0.016	0.115 ± 0.010	0.825	1.25
w/o calibration	0.225 ± 0.014	0.112 ± 0.008	0.812	1.15
w/o conformal	0.212 ± 0.012	0.058 ± 0.006	-	-
w/o positivity filter	0.218 ± 0.015	0.062 ± 0.007	0.895	1.34
PRISM (Full)	$\mathbf{0.212} \pm 0.012$	$\mathbf{0.058} \pm 0.006$	**0.908**	1.38

5.7 LLM BACKBONE SENSITIVITY

To quantify how much PRISM depends on the extractor backbone, we evaluate PRISM and key LLM baselines under multiple LLM backbones while holding the prompt templates, decoding parameters, and prompt budgets fixed. Table 3 reports the results. When PRISM uses the optional document representation E_t, the RoBERTa encoder and preprocessing are held fixed across backbone variants; thus the PRISM differences in this table primarily reflect extraction quality rather than a changing text encoder.

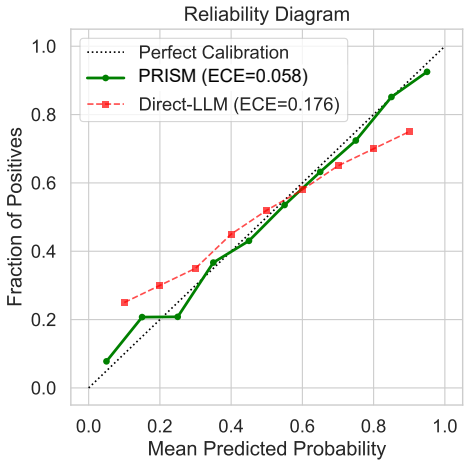

Figure 2: Reliability diagram (test set). PRISM has lower ECE than Direct-LLM (0.058 vs 0.176).

Table 1 shows that PRISM outperforms all baselines across forecast quality, calibration, and discrimination, with the largest gains in calibration (ECE drops from 0.176 for Direct-LLM to 0.058, a 67% reduction) while also improving AUROC/AUPRC. Figure 2 visualizes the calibration improvement.

Table 4 reports component ablations on reliability and uncertainty layers.

Ablation trend: Removing the Bayesian measurement layer makes probabilities overconfident and degrades both ECE and 90% rolling conformal coverage under an otherwise fixed protocol. The full PRISM conformal layer reaches pooled empirical coverage of 0.908 at the 90% target, but this reliability has an informativeness cost: in a binary label space, the average set size of 1.38 implies that roughly 38% of outputs are the non-committal $\{0, 1\}$ set rather than a singleton prediction. This is the intended reliability–informativeness trade-off, not a distribution-free guarantee under time dependence. The rolling diagnostics in Appendix C.6 show six-month coverage ranging from 86.9% to 93.5%; local dips indicate periods where nonstationarity temporarily outpaces the trailing calibration window even though aggregate coverage meets the target. Since the final prediction sets are nonempty after the fallback rule and the pre-fallback empty-set rate for full PRISM is below 0.1%, this average set size corresponds approximately to 62% singleton outputs and 38% $\{0, 1\}$ outputs. Appendix C.6 and Appendix D.5 provide additional score distributions, threshold trajectories, and coverage-by-month plots for this ablation. We report overlap/placebo sensitivity diagnostics for attribution and additional conformal set-size/coverage diagnostics in Appendix C.7 and Appendix D.5.

5.8 SEMI-SYNTHETIC VALIDATION FOR CONTRAST RECOVERY

Real-world conflict data rarely provide ground truth factor states or ground truth contrast targets. We therefore construct a semi-synthetic benchmark with a known latent factor and controlled measurement noise, then evaluate contrast recovery, absolute error and interval coverage, under standard and misspecified settings. For a selected factor index

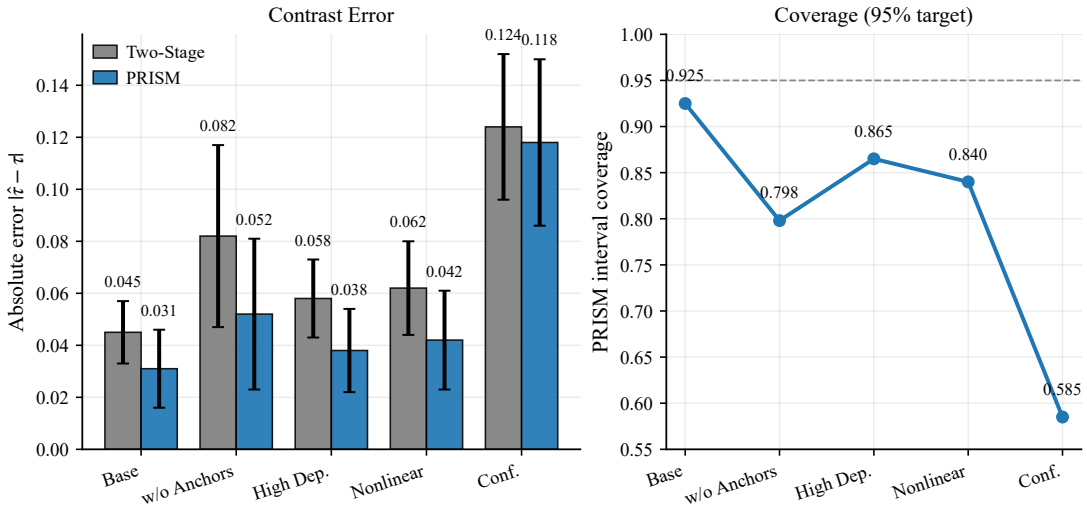

Figure 3: Semi-synthetic contrast recovery summary. Left: mean absolute contrast error for PRISM vs. Two-Stage across scenarios. Right: empirical coverage of PRISM intervals.

k, we sample

$$Z_{t,k}^{\text{syn}} \sim \text{Bernoulli}\left(\sigma\left(u^\top X_t\right)\right), \tag{11}$$

$$\tilde{Z}_{t,k}^{(m),\text{syn}} \mid Z_{t,k}^{\text{syn}} \sim \text{correlated measurement model with known } \left(\rho_k^{(1)}, \rho_k^{(0)}\right), \tag{12}$$

$$Y_{t+\Delta}^{\text{syn}} \sim \text{Bernoulli}\left(\sigma\left(\text{logit}(p_t^0) + \tau_k^{\text{true}} Z_{t,k}^{\text{syn}}\right)\right). \tag{13}$$

where p_t^0 is a baseline risk estimated from covariates and τ_k^{true} controls contrast size. Here the semi-synthetic $Z_{t,k}^{\text{syn}}$ is binary by construction; PRISM still fits a continuous intensity layer and can be interpreted as modeling the probability of factor presence under this data-generating process. We evaluate recovery under a base setting and stress tests that increase cross-prompt dependence, introduce nonlinear outcome effects, add an unobserved confounder, and add mild nonstationarity. Figure 3 summarizes the results. The coverage drop in the confounding stress test is diagnostic: it occurs when the data-generating process intentionally violates the interpretation assumptions, and it should be read as a boundary case rather than as failure of the real UCDP–GDELT conformal coverage result. Appendix A.4 reports full details.

5.9 AUDIT REPORT COMPONENTS

For each held-out country-month, PRISM can emit an audit report that records the document evidence used by the extractor, the replicate factor measurements and posterior factor summaries, the calibrated predictive probability, the conformal prediction set, and factor-contrast summaries with uncertainty and interpretation boundaries. The report is intended for post-hoc inspection and error analysis; it is not an additional validation experiment and does not by itself establish a causal interpretation. Appendix A.5 details the report fields, and Appendix A.6 specifies the manual annotation protocol used for the factor-semantics spot-check.

6 LIMITATIONS AND FUTURE WORK

PRISM depends on LLM extraction quality; missing, biased, or systematically incomplete textual evidence can degrade both predictive performance and the interpretability of factor contrast reports. Our outcome model is a generalized linear model on a compact factor inventory; while this choice supports interpretability and stable uncertainty reporting under rare events, it can miss nonlinear interactions present in complex sociopolitical systems. PRISM also does not impose an explicit temporal state-space prior on Z_t or a point-process mechanism for event dynamics in its main specification; instead it relies on text evidence at time t (and optional lagged-outcome baselines for context). Our time-adaptive conformal procedure is evaluated as an empirical coverage objective under temporal dependence rather than as a distribution-free guarantee; strong distribution shifts can still break coverage. The post-hoc calibration layer is an explicit model-mismatch correction and does not preserve a strict Bayesian posterior interpretation. Quantitative factor contrasts are reported as controlled predictive associations by default, and an intervention-style reading requires explicit identification assumptions and should be interpreted together with overlap/placebo diagnostics and sensitivity analyses. PRISM is intended as a reliability and auditability layer that is largely orthogonal to improvements in the LLM prompting and retrieval stack, so we avoid making a claim of being the strongest possible end-to-end forecaster under all prompting regimes. While we compare against

representative Direct-LLM and supervised text baselines under a fixed split protocol, stronger LLM forecasting baselines (e.g., retrieval/tool-use prompting, self-consistency with reasoning constraints, and task-specific tuning) and stronger temporal deep models (e.g., time-series transformers) are important additions for future benchmarking. Our empirical evaluation is limited to a single conflict benchmark (UCDP–GDELT). Broadening to additional event domains requires building and validating domain-specific factor inventories, ensuring split-respecting data construction, and paying higher extraction costs. We leave such cross-domain benchmarking to future work. Future work includes dynamic latent-factor priors and point-process extensions, richer measurement models for non-binary and structured factors, interference-aware estimands for geopolitical settings, and extensions to multi-event and multi-horizon forecasting.

7 CONCLUSION

PRISM reframes LLM-based event prediction by separating semantic extraction from statistical decision making. Under explicit assumptions, it treats LLM outputs as noisy measurements of latent factors, performs Bayesian posterior predictive inference that propagates extraction and measurement uncertainty into predictive distributions, and adds low-capacity calibration and time-adaptive conformal prediction for empirical reliability. PRISM further supports auditability by reporting grounded factor measurements and quantitative factor-contrast summaries with uncertainty and stated interpretation boundaries.

Acknowledgements

This work is supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123301). This work is supported by the National Natural Science Foundation of China (No. 62576340, No. U24A20335). This work is also sponsored by the Beijing Nova Program (20250484750) and the Youth Innovation Promotion Association CAS.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2022. URL https://arxiv.org/abs/2202.13415.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*, 2024.

Noah Hollmann, Samuel Müller, Leon Purucker, Arun Krishnakumar, Maximilian Körfer, S. B. Hoo, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Elvis Hsieh, Preston Fu, and Jonathan Chen. Reasoning and tools for human-level forecasting, 2024. URL https://arxiv.org/abs/2408.12036.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. Forecastbench: A dynamic benchmark of AI forecasting capabilities, 2024. URL https://arxiv.org/abs/2409.19839.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2796–2804, 2018.

Kalev Leetaru and Philip A. Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. *ISA Annual Convention*, 2013.

Yuan Li, Benjamin I. P. Rubinstein, and Trevor Cohn. Truth inference at scale: A bayesian model for adjudicating highly redundant crowd annotations, 2019. URL https://arxiv.org/abs/1902.08918.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework, 2024. URL https://arxiv.org/abs/2412.07031. arXiv preprint (v1 Dec 2024; revised v4 Dec 2025).

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. Reasoning over uncertain text by generative large language models, 2024. URL https://arxiv.org/abs/2402.09614.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alex J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1357–1370. Association for Computational Linguistics, March 2024. doi: 10.18653/v1/2024.findings-eacl.91. URL https://aclanthology.org/2024.findings-eacl.91/.

Jakob G. Rasmussen. Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512, 1986.

Philipp Schoenegger and Peter S. Park. Large language model prediction capabilities: Evidence from a real-world forecasting tournament, 2023. URL https://arxiv.org/abs/2310.13014.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy, 2024. URL https://arxiv.org/abs/2402.19379.

Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. Dynamic bayesian combination of multiple imperfect classifiers, 2012. URL https://arxiv.org/abs/1206.1831.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Ralph Sundberg and Erik Melander. Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532, 2013.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12229–12272, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.728. URL https://aclanthology.org/2024.findings-acl.728/.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency, 2025. URL https://arxiv.org/abs/2502.04964.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Xingyu Wu, Kui Yu, Jibin Wu, and Kay Chen Tan. Llm cannot discover causality, and should be restricted to non-decisional support in causal discovery, 2025. URL https://arxiv.org/abs/2506.00844.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR, 2021.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.

Yichi Zhang and Ignacio Martinez. From stochasticity to signal: A bayesian latent state model for reliable measurement with llms, 2025. URL https://arxiv.org/abs/2510.23874.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. Causal graph discovery with retrieval-augmented generation based large language models, 2024. URL https://arxiv.org/abs/2402.15301.

PRISM: Calibrated Bayesian Fusion and Auditable Attribution for Reliable LLM Event Prediction from Text (Supplementary Material)

Chengyuan Jin[1,2] **Daojian Zeng**[4] **Kang Liu**[1,2] **Jun Zhao**[1,2] **Yubo Chen**[3,*]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[2]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]School of Information Engineering, Minzu University of China
[4]Hunan Normal University, China
[*]yubo.chen@muc.edu.cn

Appendix A reports additional sensitivity analyses and audit report components; Appendix B details the experimental protocol, factor inventory, and baselines; Appendix C collects inference, measurement, exchangeability, conformal, and attribution diagnostics; and Appendix D.1 provides reproducibility details for data construction, prompts, model settings, and implementation.

A SUPPLEMENTARY RESULTS

A.1 COUNTRY-SPECIFIC PARAMETER SENSITIVITY

Our main specification infers global measurement parameters using the stratified replicate subset and propagates measurement-parameter uncertainty by integrating downstream inference over $\phi^{(l)} \sim q(\phi)$. As a sensitivity analysis, we instead fit the full set of measurement parameters independently per country (using that country's replicate subset) and re-run the same downstream inference and evaluation protocol. This isolates whether cross-country pooling in the measurement layer materially affects forecasting performance or factor interpretability. Forecasting and calibration metrics are similar under this alternative, while uncertainty estimates are noisier for low-data countries due to fewer replicated samples for estimating rater-style error parameters.

A.2 FACTOR PROXIMITY ABLATION

To address concerns that some factors may be too close to the outcome definition, we performed a targeted ablation where we removed the "Armed mobilization" factor from the factor inventory and re-ran the full pipeline under the same protocol. In this ablation, AUROC changes from 0.821 to 0.807 and ECE changes from 0.058 to 0.064.

Table 5: Cost-effectiveness sweep. Costs are relative to the main configuration (1.0x) and are computed from the split sizes under the partial-replicate protocol. Standard deviations for AUROC/ECE are typically $\pm 0.01 - 0.02$.

Subset Size	M	Relative Cost	AUROC↑	ECE↓	Coverage (90%)↑	Set Size↓
5% (Main)	**5**	**1.00x**	**0.821 ± 0.018**	**0.058 ± 0.006**	**0.908**	**1.38**
20%	5	1.43x	0.824 ± 0.018	0.057 ± 0.006	0.904	1.36
10%	5	1.14x	0.823 ± 0.017	0.056 ± 0.006	0.903	1.37
2.5%	5	0.93x	0.815 ± 0.020	0.062 ± 0.008	0.895	1.41
1%	5	0.89x	0.804 ± 0.023	0.075 ± 0.010	0.882	1.50
5%	7	1.07x	0.820 ± 0.018	0.057 ± 0.005	0.903	1.37
5%	3	0.93x	0.817 ± 0.017	0.065 ± 0.007	0.899	1.41
5%	1	0.86x	0.788 ± 0.019	0.118 ± 0.011	0.860	1.65

A.3 COST-EFFECTIVENESS CURVE

We sweep the replicate subset size (1%, 2.5%, 5%, 10%, 20%) and replicate count M (1, 3, 5, 7) and summarize the resulting performance–cost tradeoffs in Table 5. Results show diminishing returns beyond a 5% replicate subset with $M = 5$ under this protocol. This operating point is not intended as a universal rule for new domains; in a new event domain, we recommend starting with a pilot replicated subset and increasing it until measurement-parameter uncertainty, agreement PPCs, and seed-index diagnostics stabilize.

A.4 SEMI-SYNTHETIC DATA-GENERATING PROCESS

We provide additional details for the semi-synthetic validation summarized in Section 5.8. We generate synthetic latent factors and outcomes while keeping real covariates (and optionally text representations) fixed, then run PRISM using only $\tilde{Z}_{t,k}^{(m),\mathrm{syn}}$ and covariates to estimate $\hat{\tau}_k$ and evaluate absolute error and interval coverage, with stress tests that vary dependence (κ), outcome nonlinearity, and unmeasured confounding.

Anchors and stress tests. The "w/o Anchors" row in Table 6 corresponds to the main setup (no anchor labels, using only the weak orientation prior). The "Base" row treats 5% of the synthetic $Z_{t,k}^{\mathrm{syn}}$ values as observed (anchors) to fix orientation. Stress tests perturb the data-generating process away from the fitted model class by increasing cross-prompt dependence, introducing nonlinear outcome effects, adding an unobserved confounder, and adding mild nonstationarity in p_t^0.

Table 6: Semi-synthetic validation of contrast recovery under misspecification stress tests. Means and standard deviations over 50 simulation runs. "Conf." denotes unobserved confounding.

Scenario	Two-Stage $\lvert\hat{\tau} - \tau\rvert\downarrow$	PRISM $\lvert\hat{\tau} - \tau\rvert\downarrow$	PRISM 95% CI Cov.$\uparrow$
Standard ($\tau = 0.10$)			
Base ($\tau = 0.10$)	0.045 ± 0.012	$\mathbf{0.031} \pm 0.015$	0.925
w/o Anchors	0.082 ± 0.035	0.052 ± 0.029	0.798
Stress Tests (True $\tau = 0.10$)			
High Dep. ($\kappa \uparrow$)	0.058 ± 0.015	0.038 ± 0.016	0.865
Nonlinear Outcome	0.062 ± 0.018	0.042 ± 0.019	0.840
Conf. ($U \to Z, Y$)	0.124 ± 0.028	0.118 ± 0.032	0.585

A.5 AUDIT REPORT COMPONENTS

For each held-out country-month, PRISM can produce a structured audit report under the same fixed protocol used for evaluation. The report contains:

1. document identifiers and evidence spans returned by the extractor for each factor;
2. replicate factor measurements and posterior summaries of latent factor intensities;
3. the posterior predictive probability before and after calibration;
4. the conformal prediction set at the target coverage level, together with the thresholding rule used to form it;
5. factor-contrast summaries with uncertainty intervals and explicit interpretation boundaries.

These fields are intended to make the prediction traceable to text evidence and model components, and to support post-hoc inspection of extraction errors, calibration behavior, and attribution boundaries. The audit report should be read as a transparency artifact rather than as evidence for a causal effect.

A.6 MANUAL AUDIT PROTOCOL AND ANNOTATION INSTRUCTIONS

This audit checks whether the latent-factor semantics inferred by PRISM align with the operational factor definitions when applied to real document sets. As a spot-check of factor semantics, we sample 100 document–factor pairs from the held-out test span, restricted to pairs (c, t, k) where $D_{c,t}$ is non-empty and factor k belongs to the fixed inventory. Each sampled

item therefore corresponds to a specific country-month and factor. The annotator receives the factor name and operational definition (Table 7), the country and month, and document titles/excerpts sufficient to apply the definition, but does not see PRISM outputs, prompt measurements, or the target outcome Y.

The annotator assigns `present=1` only when at least one document in $D_{c,t}$ contains evidence satisfying the operational definition at a materially meaningful level for the country-month unit; ambiguous or low-intensity evidence is labeled `present=0`. We compare this label with PRISM's posterior-mode factor state dichotomized at 0.5, $\hat{z}_{c,t,k} = \mathbb{I}\{\mathrm{mode}(Z_{c,t,k} \mid \cdot) \geq 0.5\}$. The agreement rate is 83% (83/100) (95% CI: $[75\%, 90\%]$), with Cohen's $\kappa = 0.68$; disagreements are concentrated in ambiguous low-intensity cases. The confidence interval is a standard binomial proportion interval with $n = 100$ trials.

B EXPERIMENTAL PROTOCOL

Algorithm 1 gives the end-to-end implementation order used in the experiments. Here γ is the fixed extraction prompt template, $S_{\text{replicate}}$ is the stratified subset with repeated LLM measurements, ϕ collects measurement parameters, θ collects outcome-model parameters, and $S_t^{(m)}$ denotes evidence spans returned by replicate m.

Algorithm 1 PRISM (joint modeling and inference pipeline)

Input: documents $\{D_t\}_{t=1}^{T}$, outcomes $\{Y_{t+\Delta}\}_{t=1}^{T-\Delta}$, horizon Δ, prompt template γ, time-ordered splits $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{cal}}, \mathcal{I}_{\text{test}}$
// Partial replicate design: stratified subset for error estimation
Draw stratified subset $S_{\text{replicate}} \subseteq \mathcal{I}_{\text{train}}$ (5% of training data)
for $t \in S_{\text{replicate}}$ **do**
    for $m = 1$ to M **do**
        Query LLM with D_t, template γ, and seed m to obtain $(\tilde{Z}_t^{(m)}, S_t^{(m)})$
    end for
end for
for $t \in (\mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}}) \setminus S_{\text{replicate}}$ **do**
    Query LLM with D_t, template γ, and seed 1 to obtain $(\tilde{Z}_t^{(1)}, S_t^{(1)})$
end for
Fit an approximate posterior $q(\phi)$ over global measurement parameters using $S_{\text{replicate}}$
For $l = 1$ to L: draw $\phi^{(l)} \sim q(\phi)$ and infer $p(\theta, \{Z_t\}_{t \in \mathcal{I}_{\text{train}}} \mid \tilde{Z}_{\mathcal{I}_{\text{train}}}, E_{\mathcal{I}_{\text{train}}}, Y_{\mathcal{I}_{\text{train}}}, \phi^{(l)})$ using gradient-based inference over *continuous* latent factor intensities $Z_{t,k} \in (0, 1)$ (mean-field VI for full-scale runs; NUTS on a verification subset as an audit)
For each $t \in \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}}$: compute $\hat{p}_t = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}\big[P(Y_{t+\Delta} = 1 \mid Z_t, E_t, \theta) \mid \tilde{Z}_t, E_t, \phi^{(l)}\big]$ by Monte Carlo over posterior draws; this propagates $q(\phi)$ learned on $S_{\text{replicate}}$ to non-replicated t
Fit calibration map $c(\cdot)$ on $\mathcal{I}_{\text{cal}}$ only and set $\hat{p}_t^{\text{cal}} = c(\hat{p}_t)$ for $t \in \mathcal{I}_{\text{test}}$
for $t \in \mathcal{I}_{\text{test}}$ in chronological order **do**
    Conformal: output prediction set C_t using only past labeled scores (calibration split and earlier test points with observed $Y_{i+\Delta}$), with a gap window to respect horizon Δ
end for
Attribution: estimate $\{\hat{\tau}_k\}_{k=1}^{K}$ with uncertainty via posterior averaging under the fitted outcome model (Appendix C.8)
Output: $\hat{p}_t^{\text{cal}}$ for $t \in \mathcal{I}_{\text{test}}$, conformal sets $\{C_t\}_{t \in \mathcal{I}_{\text{test}}}$, attribution report, evidence spans $\{S_t^{(m)}\}_{m \in \mathcal{M}_t}$ where $\mathcal{M}_t = \{1, \ldots, M\}$ if $t \in S_{\text{replicate}}$ else $\{1\}$

B.1 FACTOR INVENTORY USED IN MAIN EXPERIMENTS

Table 7: Factor inventory and operational definitions used in main experiments ($K = 8$).

k	Factor	Operational definition (country-month)
1	Protest	Reports of sizable protests, strikes, riots, or mass demonstrations involving domestic actors.
2	Crackdown	State-led repression or coercive measures, arrests, curfews, violent dispersal, targeting domestic opposition or civilians.
3	Armed mobilization	Reports of military buildup, troop movements, arms transfers, militia recruitment, or mobilization of armed groups *without* reported battle deaths.
4	Communal tension	Salient ethnic, sectarian, or communal tensions, hate incidents, or localized communal violence signals.
5	Economic shock	Major economic disruptions plausibly linked to instability, currency crisis, fuel or food shortages, inflation spikes, widespread unemployment shocks.
6	International pressure	Sanctions, threats of sanctions, diplomatic isolation, major condemnations, or external coercive pressure directed at the country.
7	Political transition	Elections, coups, leadership contests, major cabinet changes, or disputed transitions with credible contention signals.
8	Peace process	Ceasefire talks, peace negotiations, mediation, demobilization plans, or formal de-escalation agreements under discussion or implementation.

We draft candidate factors from domain literature and text-only exploration (e.g., unsupervised clustering and keyword mining) without using outcome labels. For each factor we specify an evidence-based inclusion/exclusion rule at the country-month unit; a factor is marked present only if at least one document contains direct textual evidence meeting the operational definition at a materially meaningful level. We reduce redundancy by merging or refining overlapping definitions using text-only checks (e.g., co-occurrence and recurring evidence patterns), and we lock the inventory and extraction prompt template before fitting and reporting results.

B.2 ADDITIONAL BASELINE SPECIFICATIONS

This appendix clarifies baseline instantiations so that differences reflect modeling choices rather than input or protocol mismatches.

Shared protocol across baselines. All methods use identical train/calibration/test splits, identical document payloads $D_{c,t}$, identical factor inventory and operational definitions when applicable (Table 7), and the same evaluation protocol. Any tuning is performed on the calibration split only; no method selection or hyperparameter choice uses test labels.

Direct-LLM. The model receives $D_{c,t}$ and a forecasting prompt that asks for a probability for $Y_{c,t+\Delta}$. We use the same number of prompt calls M and the same decoding settings as PRISM. The output probability is taken as the LLM's numeric response (after deterministic parsing). Phrasing robustness uses paraphrases of the forecasting question while keeping the same horizon, country-month, and closed-book constraint.

Direct-LLM (factor-aware). The model is prompted to (i) answer each factor question using the operational definition and evidence quotes from $D_{c,t}$, and then (ii) output a forecast probability conditioned on those extracted factor states. This controls for providing factor structure while keeping the probability generation inside the LLM.

LLM-Ensemble. We average probabilities from M Direct-LLM prompt variants and seeds, optionally followed by the same calibration and conformal procedures used for PRISM.

Text-only predictor. We train a supervised classifier on document representations $E_{c,t}$ (constructed from $D_{c,t}$) to predict $Y_{c,t+\Delta}$. The classifier is trained on the training split and any calibration is learned on the calibration split. Conformal prediction, when used, follows the same gapped rolling protocol as PRISM.

Text+time classical baselines. We include time-series logistic regression with lagged outcomes and controls, and gradient-boosted trees on document features and controls. All lags are computed strictly from past months and aligned to the same country-month grid.

Point-process baselines. For context from the event modeling literature, we include a Hawkes-style baseline that models event intensities using past event history [Rasmussen, 2013]. Since our main benchmark uses country-month binary labels rather than event timestamps, we implement a discretized intensity model on the same grid and evaluate it under the same evaluation protocol. We treat this as a history-driven baseline family that complements text-driven predictors rather than as a direct substitute for text-based extraction.

Two-stage measurement+prediction. We first infer latent factor posteriors from prompt measurements using a Dawid–Skene-style model (with the same orientation conventions as PRISM) and then fit an outcome model using posterior means (or posterior draws) of Z_t as inputs. This isolates the value of *joint* uncertainty propagation in PRISM versus a plug-in approach.

LLM-driven Bayesian inference. When feasible, these baselines use an LLM to propose a probabilistic model specification (e.g., a PPL program) from a fixed template, and then fit that model using the same train/calibration splits. We treat this as a model-construction baseline rather than as an additional source of future information; prompts are restricted to use only the provided data schema and do not include any test labels.

C DIAGNOSTICS, METRICS, AND ATTRIBUTION

C.1 INFERENCE AND IDENTIFIABILITY AUDIT DETAILS

This appendix specifies the minimum diagnostic set we use to reduce the risk that approximate inference or label orientation conventions distort uncertainty reporting or factor semantics.

Mean-field VI vs NUTS: beyond posterior means. On the NUTS verification subset, we compare VI and NUTS on posterior means of $\hat{p}_t$ over matched time points, posterior quantiles and interval widths for $\hat{p}_t$, posterior summaries (including interval widths) for key global parameters such as β_k and calibration-map parameters when used, and posterior summaries for factor-contrast summaries τ_k and their interval widths. We treat agreement of interval widths and quantiles as a necessary condition for using VI-derived intervals for interpretation; when VI produces systematically narrower intervals, we report contrast and uncertainty claims conservatively and rely on conformal evaluation for predictive uncertainty sets.

Interval coverage checks on semi-synthetic ground truth. Since VI is known to underestimate posterior variance under mean-field assumptions, we additionally evaluate *frequentist coverage* of posterior credible intervals on semi-synthetic data where the ground truth is known (Appendix A.4). For each simulated replicate, we fit the model using both VI and NUTS (at reduced scale) and record coverage of nominal 90% and 95% intervals for: (i) predictive probabilities $\hat{p}_t$ at held-out points, (ii) key outcome coefficients β_k, and (iii) contrast summaries τ_k. We treat systematic under-coverage of VI intervals as evidence that mean-field is too restrictive for uncertainty reporting at scale, and in that case we (a) switch to a richer variational family (e.g., full-rank covariance on the outcome block and structured dependence for (Z, θ)) for reporting, and (b) report conformal sets as the primary uncertainty interface for decisions.

Propagation of measurement-parameter uncertainty. Because downstream uncertainty depends on both latent-factor uncertainty and measurement-parameter uncertainty, we audit the effect of integrating over ϕ by comparing posterior predictive intervals and contrast-interval widths under two variants: (i) plug-in $\phi = \mathbb{E}_{q(\phi)}[\phi]$ and (ii) Monte Carlo integration $\phi^{(l)} \sim q(\phi)$. When the integrated variant substantially widens intervals or changes ranking of uncertainty, we treat plug-in as potentially overconfident and default to the integrated variant for all uncertainty reporting.

Orientation and label-switching diagnostics. For each factor k, we summarize posterior evidence that the fitted branch is consistent with the intended orientation convention: we check whether error-rate parameters concentrate away from the chance region (e.g., $\rho_k^{(1)}$ and $\rho_k^{(0)}$ away from 0.5), whether orientations are stable across multiple random initializations on the verification subset, and whether agreement-based posterior predictive checks (pairwise prompt agreement) indicate measurement-layer misspecification that could mask label ambiguity. Concretely, we run multi-start inference and align each run to a reference solution by applying the orientation relabeling (flip/no-flip) that maximizes agreement of posterior mean

latent intensities on the replicate subset; if multiple aligned modes remain with comparable fit, or if aligned orientations disagree across runs, we treat the factor as weakly identified. When a factor exhibits weak signal or posterior ambiguity under this audit, we treat its semantic interpretation as low-confidence and avoid strong causal wording for that factor's attribution report; when anchor labels are available in semi-synthetic settings, we additionally verify that the aligned orientation matches the anchor direction.

Table 8: Posterior predictive checks (PPC) for measurement agreement. We report the posterior predictive p-value for the observed pairwise agreement rate $\hat{a}_k$ under the fitted model. Values near 0 or 1 indicate potential misfit (systematic disagreement or over-agreement not captured by κ_k). Most factors fall within the non-extreme range; "Political transition" is marginal (rounded $p = 0.04$).

Factor	Observed Agreement $\hat{a}_k$	PPC p-value
1. Protest	0.82	0.09
2. Crackdown	0.91	0.65
3. Armed mobilization	0.76	0.38
4. Communal tension	0.85	0.51
5. Economic shock	0.92	0.88
6. International pressure	0.89	0.68
7. Political transition	0.79	0.04
8. Peace process	0.88	0.55

C.2 MEASUREMENT DEPENDENCE AND AGREEMENT CHECKS

Repeated LLM queries are not guaranteed to be conditionally independent given the latent factor intensity, because prompt paraphrases can share document-conditioned biases. We therefore include an exchangeable correlated-measurement extension with a factor-specific dependence parameter κ_k and validate adequacy using agreement-based posterior predictive checks on the replicate subset.

Observed agreement statistics. For each factor k and each replicated country-month (c, t), let $\tilde{Z}_{c,t,k}^{(1:M)}$ denote the M binary prompt measurements. We compute (i) the pairwise agreement rate

$$\hat{a}_{c,t,k} = \frac{2}{M(M-1)} \sum_{m<m'} \mathbb{I}\{\tilde{Z}_{c,t,k}^{(m)} = \tilde{Z}_{c,t,k}^{(m')}\}, \tag{14}$$

and (ii) the replicate-positive count $s_{c,t,k} = \sum_{m=1}^{M} \tilde{Z}_{c,t,k}^{(m)}$. We summarize the empirical distribution of $\hat{a}_{c,t,k}$ and $s_{c,t,k}$ across replicated items.

Posterior predictive check. Given posterior draws of $(Z_{c,t,k}, \kappa_k, \rho_k^{(1)}, \rho_k^{(0)})$ from the fitted measurement layer, we simulate replicated measurements $\tilde{Z}_{c,t,k}^{(1:M),\mathrm{rep}}$ and compute the corresponding replicated agreement statistics $(\hat{a}_{c,t,k}^{\mathrm{rep}}, s_{c,t,k}^{\mathrm{rep}})$. We treat systematic tail behavior of these summaries under the posterior predictive distribution as a misspecification signal at this diagnostic resolution. Table 8 reports the summary.

C.3 EXCHANGEABILITY DIAGNOSTICS FOR PROMPT REPLICATES

PRISM treats replicate measurements as conditionally exchangeable given (Z, ϕ); in practice, exchangeability can fail if replicate index (seed) induces systematic behavior. We therefore report diagnostics on the replicate subset:

Seed-index marginal checks. We report the marginal positive rates for each replicate seed index $m \in \{1, \ldots, 5\}$ across the replicate subset in Table 9. Across the reported factors, we do not observe strong evidence of seed-index differences in these marginals under a simple homogeneity test, and pairwise agreement rates between distinct seeds do not show an obvious systematic structure (e.g., Seed 1 does not consistently agree more with Seed 2 than with Seed 3).

Table 9: Seed-index diagnostic summary on the replicate subset. Marginal positive rates (MPR) are similar across the 5 fixed random seeds used for extraction, consistent with exchangeability at the level of these seed-wise marginals.

Factor	Seed 1 MPR	Seed 2 MPR	Seed 3 MPR	Seed 4 MPR	Seed 5 MPR	p-value (χ^2)
1. Protest	0.121	0.129	0.118	0.125	0.122	0.21
2. Crackdown	0.082	0.091	0.085	0.089	0.083	0.15
3. Armed mobilization	0.148	0.156	0.152	0.150	0.154	0.42
4. Communal tension	0.108	0.115	0.106	0.112	0.110	0.11
5. Economic shock	0.040	0.044	0.042	0.041	0.043	0.35
6. International pressure	0.088	0.095	0.091	0.090	0.089	0.21
7. Political transition	0.073	0.079	0.075	0.072	0.076	0.04
8. Peace process	0.036	0.041	0.038	0.037	0.039	0.28

Permutation invariance checks. Because an exchangeable model is invariant to relabelings of the replicate index, we check that posterior summaries and PPC agreement statistics are stable under random permutations of replicate labels m within each (c, t) on the replicate subset. Large changes indicate that replicate-specific structure is present and that a non-exchangeable extension may be needed.

Robustness action when violated. When a factor exhibits clear seed-index effects, we treat its measurement layer as misspecified for uncertainty reporting: we avoid fine-grained claims about ϕ for that factor, mark its attribution semantics as low-confidence, and prioritize downstream empirical reliability evaluation (calibration and conformal) for decisions. In our reported diagnostics, "Political transition" is the marginal factor under both the agreement PPC and seed-index checks (rounded $p = 0.04$), so any single-factor attribution for that factor should be read as flagged rather than as a strong semantic conclusion. A natural follow-up sensitivity check is a robustness fit with replicate-specific intercepts, which relaxes exchangeability by allowing a seed-dependent bias in the measurement likelihood.

C.4 METRIC AGGREGATION: MICRO VS MACRO

Let $\mathcal{I}$ denote the set of country-month instances in the test set. For a scoring rule or metric $M(\cdot)$ computed from predicted probabilities and labels, we define:

Micro aggregation (main tables). Pool all instances and compute once:

$$M_{\text{micro}} = M\Big(\{(\hat{p}_i, y_i)\}_{i\in\mathcal{I}}\Big). \tag{15}$$

Macro aggregation (robustness check). Compute the metric within each country and average across countries:

$$M_{\text{macro}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} M\Big(\{(\hat{p}_{c,t}, y_{c,t})\}_{t:(c,t)\in\mathcal{I}}\Big), \tag{16}$$

where $\mathcal{C}$ are countries appearing in the test set.

C.5 METRIC DEFINITIONS

We use binned expected calibration error with equal-width bins. Let $I_b = \big(\frac{b-1}{B}, \frac{b}{B}\big]$ partition $[0, 1]$, and let n_b be the number of predictions with $\hat{p} \in I_b$. Define $\text{acc}(b) = \frac{1}{n_b}\sum_{i:\hat{p}_i\in I_b} \mathbb{I}\{Y_i = 1\}$ and $\text{conf}(b) = \frac{1}{n_b}\sum_{i:\hat{p}_i\in I_b} \hat{p}_i$. Then

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{n}\,\big|\text{acc}(b) - \text{conf}(b)\big|. \tag{17}$$

Reliability diagrams use the same binning scheme.

Class-conditional calibration. To avoid majority-class masking under low base rates, we also report class-conditional ECE by restricting the above computation to indices with $Y_i = 1$ (positive class) and to indices with $Y_i = 0$ (negative class), using the same binning on $\hat{p}$.

Precision–recall and AUPRC. We compute the precision–recall curve by sweeping a threshold $\tau \in [0, 1]$ over predicted probabilities and reporting precision and recall for predicting $Y = 1$ at each τ. AUPRC is computed as the area under this curve using the standard stepwise interpolation induced by sorting predictions by $\hat{p}$.

C.6 ROLLING COVERAGE COMPUTATION FOR TIME-ADAPTIVE CONFORMAL

For each test time t, conformal outputs a prediction set $C_t \subseteq \{0, 1\}$. Define the instantaneous coverage indicator

$$\mathrm{cov}_t = \mathbb{I}\{Y_{t+\Delta} \in C_t\}. \tag{18}$$

To summarize temporal stability, we compute a rolling average over the already-observed portion of the test sequence:

$$\overline{\mathrm{cov}}_t = \frac{1}{R} \sum_{j=t-R+1}^{t} \mathrm{cov}_j, \tag{19}$$

where R is a fixed trailing window length (e.g., 6 months) chosen before test evaluation. The reported stability range (e.g., $[0.88, 0.93]$) is computed as $\min_t \overline{\mathrm{cov}}_t$ and $\max_t \overline{\mathrm{cov}}_t$ over the test months, after the initial burn-in where fewer than R test points are available. This rolling summary is distinct from the conformal calibration window of size W used to compute thresholds; it is only a diagnostic over the test trajectory and never feeds back into prediction. On the UCDP–GDELT test period, the 90% conformal target yields mean empirical coverage 90.8%, with six-month rolling coverage ranging from 86.9% to 93.5% and pooled average prediction-set size 1.38.

C.7 ATTRIBUTION DIAGNOSTICS AND SUPPORT CHECKS

This appendix specifies the diagnostic summaries reported alongside factor-contrast summaries to clarify interpretation boundaries.

Figure 4 summarizes the overlap and placebo diagnostics used to contextualize reported factor-contrast associations.

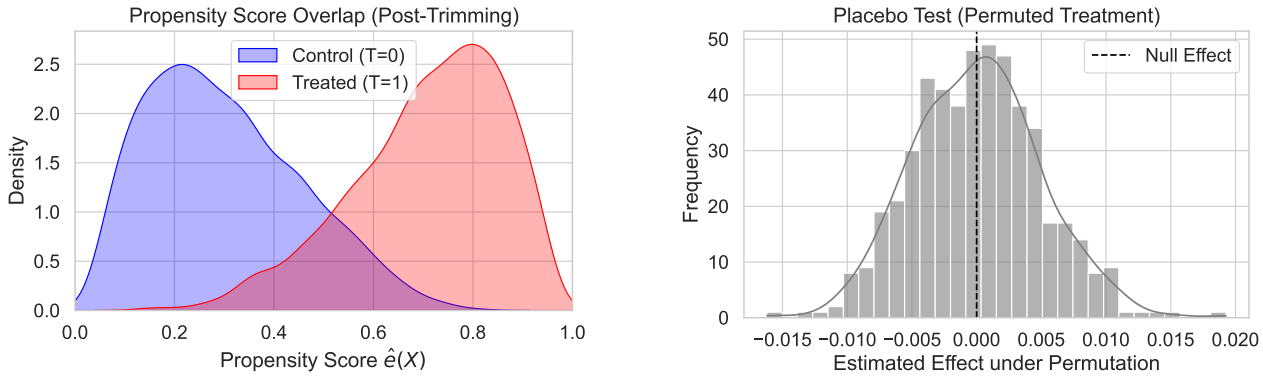

Figure 4: Attribution diagnostics for factor-contrast interpretation. Left: propensity score overlap under the configured auxiliary diagnostic model (post-trimming): extreme-mass rates $P(\hat{e} < 0.05) = 1.85\%$ for $T = 0$ and $P(\hat{e} > 0.95) = 1.89\%$ for $T = 1$ (other tails ≈ 0). Right: placebo test permuting T within time blocks to preserve marginal dependence (500 permutations); the two-sided exceedance rate $P(|\hat{\tau}_{\mathrm{perm}}| \geq 0.01) = 4.6\%$. These diagnostics contextualize controlled predictive associations and are not evidence of a causal effect.

Overlap and positivity. We report the distribution of estimated propensities computed from an *auxiliary diagnostic model* that predicts a binarized treatment indicator $\mathbb{I}\{T_t \geq q_{0.5}(T)\}$ from X_t using posterior draws (not used by PRISM for forecasting or reporting τ_k). We summarize extreme-mass rates (e.g., fraction with propensity below a small threshold or above a large threshold), together with the impact of any positivity filtering on sample size and contrast stability.

Placebo tests. We report the distribution of placebo contrast estimates under permuted T (within time blocks to preserve marginal dependence), including a two-sided exceedance summary under a fixed magnitude threshold (e.g., $|\hat{\tau}| \geq 0.01$). In Figure 4, the corresponding placebo exceedance rate is 4.6% over 500 permutations.

E-values. We report E-values computed from the posterior mean contrast mapped to a risk-ratio scale when needed, and E-values corresponding to the endpoints of the reported credible interval, so robustness can be read together with posterior uncertainty.

Support and extrapolation checks for latent treatments Because $T = Z_{t,k}$ is continuous and latent, an extreme contrast $T = 1$ vs $T = 0$ can be an extrapolation when posterior mass concentrates away from the boundaries. For each factor, we therefore summarize the empirical support of T using posterior summaries of T over time points (e.g., marginal quantiles of T) and report an in-support contrast:

$$\tau_k^{\mathrm{Q}} = \mathbb{E}\Big[\mathbb{E}[Y \mid T = q_{0.9}(T), X] - \mathbb{E}[Y \mid T = q_{0.1}(T), X]\Big], \tag{20}$$

where $q_{0.9}(T)$ and $q_{0.1}(T)$ denote upper and lower quantiles of T under the fitted posterior used for reporting. This contrast is designed to reduce reliance on regions with little posterior support while keeping the interpretation as a standardized contrast summary under the fitted outcome model.

C.8 CONTRAST COMPUTATION UNDER THE LATENT-FACTOR MODEL

We compute factor contrasts by posterior averaging under the fitted outcome model. Writing $T_t = Z_{t,k}$ and $X_t = (Z_{t,-k}, E_t)$, the primary reported estimand is the in-support quantile contrast

$$\tau_k^{\mathrm{Q}} = \mathbb{E}\Big[\mathbb{E}[Y \mid T = q_{0.9}(T), X] - \mathbb{E}[Y \mid T = q_{0.1}(T), X]\Big], \tag{21}$$

interpreted as a controlled predictive association unless additional identification assumptions are explicitly adopted. We also report the standardized extreme contrast

$$\tau_k = \mathbb{E}\Big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\Big], \tag{22}$$

as a secondary absent-versus-fully-present summary. Given posterior samples $\{Z_{1:T}^{(s)}, \theta^{(s)}\}_{s=1}^{S}$ from the joint model, we compute $m_{\theta^{(s)}}(t, x) = P(Y_{t+\Delta} = 1 \mid T = t, X = x, \theta^{(s)})$ and estimate the extreme contrast as

$$\hat{\tau}_k^{(s)} = \frac{1}{N} \sum_{t=1}^{N} \Big(m_{\theta^{(s)}}(1, X_t^{(s)}) - m_{\theta^{(s)}}(0, X_t^{(s)}) \Big), \tag{23}$$

where $X_t^{(s)}$ uses the corresponding latent-factor draw. The quantile contrast is computed analogously by replacing $(1, 0)$ with the posterior support quantiles $(q_{0.9}^{(s)}(T), q_{0.1}^{(s)}(T))$. We report posterior means and credible intervals from posterior quantiles of the corresponding sampled contrasts.

D REPRODUCIBILITY AND IMPLEMENTATION DETAILS

D.1 REPRODUCIBILITY DETAILS: DATA PROCESSING AND PROMPT SETS

Country-month construction and document filtering/deduplication For each GDELT-linked article, we extract its publication date, associated country code(s), and URL. We restrict to English articles and map each document to one or more country-month bins (c, t) using GDELT country tags and the document timestamp truncated to month. Within each (c, t) bin, we remove exact duplicates by URL, and remove near-duplicates by normalizing the title (lowercasing and stripping punctuation/whitespace) and dropping repeated normalized titles. The remaining documents are concatenated in chronological order to form $D_{c,t}$.

Outcome label construction We use UCDP GED state-based violence events and aggregate them to country-month units. For horizon $\Delta = 1$, the label is $Y_{c,t+1} = 1$ if at least one qualifying event occurs in country c during month $t + 1$ and 0 otherwise. All alignment uses calendar months; documents from month t are used to forecast month $t + 1$.

Control alignment Controls $X_{c,t}$ are aligned to monthly units by carry-forward within year using the most recent available annual value for that country. Countries are filtered if controls are missing for more than 20% of months.

Partial replicate design details To reduce redundant LLM invocations while preserving principled uncertainty quantification, we adopt a standard partial replicate design. Here we provide complete implementation details: From the training split, we draw a stratified random sample of 5% of country-months, stratifying by country identity, event base-rate tertile, and document-volume tertile. For each sampled country-month, we perform $M = 5$ repeated extractions using the single-invocation template and distinct random seeds; for all remaining samples (training remainder, calibration, test) we perform a single extraction ($m = 1$). Using only the repeated measurements from this subset, we fit the measurement model to infer an approximate posterior $q(\phi)$ over factor-specific global parameters (sensitivity $\rho_k^{(1)}$, specificity $\rho_k^{(0)}$, and cross-replicate correlation κ_k). Downstream forecasting integrates over $\phi^{(l)} \sim q(\phi)$, and we also report a plug-in variant using $\phi = \mathbb{E}_{q(\phi)}[\phi]$ as an ablation for computational comparison. Sensitivity analysis compares 1%, 2.5%, 5%, 10%, and 20% replicate subsets; performance changes little beyond the 5% operating point but degrades more clearly at 1%.

Repeated measurements via fixed random seeds ($M = 5$) All replicates use the *same* prompt template shown in the main text, which asks the model to assess all K factors in a single call. To generate $M = 5$ exchangeable replicates without manual prompt engineering, we use fixed distinct random seeds. The full prompt template is:

```
System: You are an information extraction system. Use only the provided
documents.
User: Country = {c}. Month = {t}.
Input documents: a list of items {doc_id, text}. Each text is news content for
this country-month.
Task: For each factor in the inventory, decide whether it is materially
present in this country-month based only on explicit evidence in the
documents.
Factor inventory: {(FACTOR_k, OPERATIONAL_DEFINITION_k)}_{k=1}^{K}.
Rules:
(i) If evidence is insufficient, set present=NO and confidence=0.5.
(ii) Evidence must be 1-3 verbatim quotes copied from text and prefixed with
the corresponding doc_id.
(iii) Do not use world knowledge; do not infer beyond the text.
Return only valid JSON (no markdown, no extra keys) with keys:
{
  "country": "{c}",
  "month": "{t}",
  "factors": [
    {
      "name": "...",
      "present": "YES" or "NO",
      "confidence": 0.0 to 1.0,
      "evidence": ["DOC_ID: quote", "..."]
    },
    ...
  ]
}
```

We generate $M = 5$ replicates using deterministic seeds derived from (c, t, m) for $m \in \{1, \ldots, 5\}$. All replicates use the same decoding parameters (temperature $= 0.2$, top-$p = 0.95$, max tokens $= 256$). This reduces confounding from manual prompt engineering and supports reproducibility while preserving exchangeability of the replicates.

Experiment configuration disclosure We fix the prompt budget M, decoding parameters, factor inventory, and data splits before evaluation. We choose calibration and conformal hyperparameters on the calibration split using a small fixed candidate set: calibration is either *off* or isotonic; conformal uses a fixed target level ($1 - \alpha = 0.90$) and selects window and gap parameters from a small preset grid by minimizing calibration-split empirical miscoverage and average set size.

For reproducibility, we record the random seeds used for extraction replicates, inference initialization, and any stochastic training components in non-Bayesian baselines.

D.2 ADDITIONAL DISCLOSURE FOR REPRODUCIBILITY

LLM Model and Cutoff Details For all main experiments, we use: **Model:** Llama-3-70B-Instruct (Meta, 2024); **knowledge cutoff:** provider-documented; **serving:** vLLM; **decoding parameters:** temperature $= 0.2$, top-$p = 0.95$, max tokens $= 256$ (as reported in main text).

RoBERTa Text Representation Details For text-only baselines and optional E_t covariates: We use RoBERTa-Large as the encoder. For each document set $D_{c,t}$, we embed each document independently using a first-token pooled representation and then take the mean across documents in $D_{c,t}$ to form $E_{c,t}$. Documents are truncated to 512 tokens before embedding. When used alongside factors Z_t, we compute residualized embeddings $E_t^{\perp}$ by regressing E_t on Z_t on the training split and retaining the residuals.

DeBERTa-v3 supervised text model. For the supervised text baseline in Table 1, we use a DeBERTa-v3 encoder as the document embedder in place of RoBERTa, form $E_{c,t}$ using the same per-document embedding and mean aggregation described above, and train a supervised classifier on top of $E_{c,t}$ using only the training split with hyperparameters selected on the calibration split under the same split protocol.

D.3 CONFORMAL PREDICTION NONCONFORMITY AND IMPLEMENTATION

Full conformal implementation details: We use $s_t(y) = 1 - \hat{p}_t^{\mathrm{cal}}(y)$, where $\hat{p}_t^{\mathrm{cal}}(1) = \hat{p}_t^{\mathrm{cal}}$ and $\hat{p}_t^{\mathrm{cal}}(0) = 1 - \hat{p}_t^{\mathrm{cal}}$. For each label $y \in \{0, 1\}$, we compute $q_{t,y}$ as the $(1 - \alpha)$ quantile of $\{s_i(y) : i \in [t - W - G, t - G - 1], Y_i = y\}$. If $n_{t,y} < n_{\min} = 5$, including during the positive-class warm-up period in this rare-event benchmark, we use the marginal threshold $q_t = \mathrm{Quantile}_{1-\alpha}(\{s_i(Y_i) : i \in [t - W - G, t - G - 1]\})$ for that label. The prediction set is $C_t = \{y \in \{0, 1\} : s_t(y) \leq q_{t,y}\}$, and if $C_t = \emptyset$ we set $C_t = \{0, 1\}$. When reporting diagnostics, we record whether $C_t = \emptyset$ *before* applying the $\{0, 1\}$ fallback replacement.

D.4 METHOD-COMPARISON UNCERTAINTY PROTOCOL

We quantify uncertainty for method differences using country-level paired comparisons and nonparametric uncertainty intervals: on the test set, we compute each metric (NLL, Brier, ECE, AUROC) separately for each country using that country's country-month instances, then compute per-country paired differences (PRISM minus comparator). We form 95% intervals by resampling countries with replacement and recomputing the mean paired difference, preserving each country's within-country dependence while using countries as the resampling units. We report effect sizes (mean paired differences) together with these intervals and interpret them descriptively given dependence and heterogeneity. Unless otherwise stated, we use $B_{\mathrm{boot}} = 200$ country-level resamples.

D.5 CONFORMAL ABLATION DIAGNOSTICS

For the "w/o measurement" ablation, Table 10 summarizes calibration and conformal dynamics.

Table 10: Diagnostics for the "w/o Measurement" ablation. Mean Confidence denotes the micro-averaged calibrated probability $\mathbb{E}[\hat{p}^{\text{cal}}]$ over instances. "Pre-fallback Empty Set Rate" records the fraction of times $C_t = \emptyset$ before the $\{0, 1\}$ fallback replacement.

Metric	PRISM (Main)	w/o Measurement	Note
Score Calibration			
Mean Confidence	0.081	0.245	Overconfident
Empirical Event Rate	0.079	0.079	
ECE	0.058	0.115	Miscalibrated
Conformal Dynamics			
Mean Threshold $q_{0.9}$	0.22	0.04	Too strict
Threshold Volatility (std)	0.04	0.12	Unstable
Pre-fallback Empty Set Rate ($C_t = \emptyset$)	<0.1%	25%	Before $\{0, 1\}$ fallback

E ADDITIONAL MODELING AND EXTENSIONS

E.1 NON-BINARY FACTORS: ORDINAL AND CONTINUOUS EXTENSIONS

The main text presents the binary-factor instantiation. PRISM extends to ordinal or continuous factor measurements by keeping the same joint posterior predictive semantics while changing the likelihood links used in the measurement and/or outcome components.

Ordinal factor measurements. Suppose factor k is measured on an ordered scale $\tilde{Z}_{t,k}^{(m)} \in \{1, \dots, L\}$ (e.g., none / low / medium / high). We retain a continuous latent intensity $Z_{t,k} \in (0, 1)$ and define an ordered-logistic measurement likelihood with cutpoints $b_{k,1} < \cdots < b_{k,L-1}$ and prompt-specific noise scale $s_k > 0$:

$$P(\tilde{Z}_{t,k}^{(m)} \le \ell \mid Z_{t,k}) = \sigma\left(\frac{b_{k,\ell} - \text{logit}(Z_{t,k})}{s_k}\right), \quad \ell = 1, \dots, L-1, \tag{24}$$

with the implied category probabilities given by adjacent differences. This preserves the interpretation of $Z_{t,k}$ as a latent intensity while allowing ordinal reported measurements. The correlated-measurement layer can be retained by introducing a latent per-(t, k) propensity and treating the ordered-logistic probabilities as the mean parameters for an overdispersed random-effect variant; in practice we use the simpler conditionally independent ordinal likelihood when ordinal labels are available.

Continuous factor measurements. For continuous measurements (e.g., a factor-specific intensity score), we use a Gaussian (or Student-t) measurement likelihood:

$$\tilde{Z}_{t,k}^{(m)} \mid Z_{t,k} \sim \mathcal{N}(g_k(Z_{t,k}), \sigma_k^2), \tag{25}$$

where $g_k(\cdot)$ maps $(0, 1)$ to the measurement scale (often $g_k(Z) = \text{logit}(Z)$ or $g_k(Z) = Z$ if the measurement is already in $[0, 1]$). A Student-t likelihood is a drop-in replacement when heavy-tailed measurement noise is observed.

Outcome link choice. When $Z_{t,k}$ is non-binary, the outcome model can use a linear effect on the log-odds, $\sigma(\alpha + \beta^\top Z_t + w^\top E_t)$, or a monotone transformation (e.g., spline basis) when domain knowledge suggests nonlinearity. The posterior predictive probability remains

$$\hat{p}_t = \int P(Y_{t+\Delta} = 1 \mid Z_t, E_t, \theta)\, p(Z_t, \theta \mid \tilde{Z}_{1:t}, E_{1:t}, Y_{1:t-\Delta})\, dZ_t\, d\theta, \tag{26}$$

so uncertainty from non-binary measurements propagates identically to the binary case.

E.2 PRIOR SPECIFICATION AND SENSITIVITY SETTINGS

Table 11 summarizes the core definitions and priors used in the main specification.

Table 11: Core definitions and priors used in the main specification.

Symbol	Role	Main prior / setting
$Z_{t,k}$	latent factor intensity	$Z_{t,k} \mid \pi_k, \tau_z \sim \mathrm{Beta}(\tau_z \pi_k, \tau_z(1-\pi_k))$ with $\tau_z = 2$ and $\pi_k \sim \mathrm{Beta}(1,1)$
$\tilde{Z}_{t,k}^{(m)}$	LLM measurement replicate	$\tilde{Z}_{t,k}^{(m)} \mid Z_{t,k} \sim \mathrm{Bernoulli}(\rho_k^{(1)} Z_{t,k} + (1-\rho_k^{(0)})(1-Z_{t,k}))$
$\rho_k^{(1)}, \rho_k^{(0)}$	sensitivity / specificity	weakly informative Beta priors; estimated on a replicated subset and propagated downstream (Appendix D.1)
κ_k	cross-replicate dependence	dependence-aware extension with agreement checks (Appendix C.2)
$\theta = (\alpha, \beta, w)$	outcome parameters	$\alpha \sim \mathcal{N}(0, 2.5^2)$; $\beta, w \sim$ Student-$t(3, 0, 2.5)$
K, M	factors / replicates	$K = 8$, $M = 5$ in main runs (Table 2)

For the outcome model parameters $\theta = (\alpha, \beta, w)$ in the logistic link, we use weakly regularizing priors on the log-odds scale:

$$\alpha \sim \mathcal{N}(0, 2.5^2), \tag{27}$$

$$\beta_j \sim \text{Student-}t(\nu = 3, 0, 2.5), \quad j = 1, \ldots, K, \tag{28}$$

$$w_\ell \sim \text{Student-}t(\nu = 3, 0, 2.5), \quad \ell = 1, \ldots, \dim(E). \tag{29}$$

For latent factor intensities, we use

$$Z_{t,k} \mid \pi_k, \tau_z \sim \mathrm{Beta}\big(\tau_z \pi_k, \tau_z(1-\pi_k)\big), \quad \pi_k \sim \mathrm{Beta}(1,1), \tag{30}$$

and we set $\tau_z = 2$ in the main specification to keep the intensity prior weak while avoiding extreme mass at $0/1$ that can interact with rare-event outcomes. Sensitivity analysis varies (i) the prior scales for (α, β, w), (ii) measurement-parameter priors (including $(\rho_k^{(1)}, \rho_k^{(0)})$ and κ_k), and (iii) the latent-intensity concentration $\tau_z \in \{1, 2, 5, 10\}$, and reports the resulting changes in forecasting, calibration, and contrast summaries together with posterior predictive checks. Figure 5 visualizes the performance sensitivity summary.

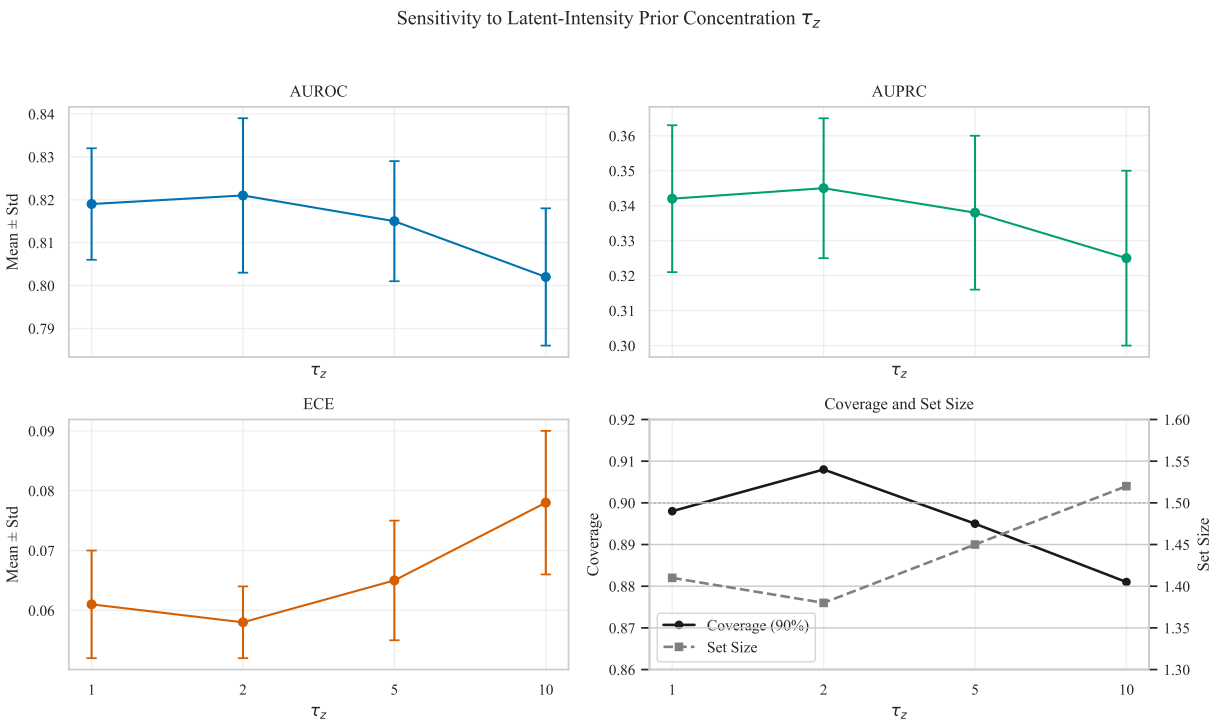

Figure 5: Sensitivity to the latent-factor prior concentration τ_z. We plot forecasting and calibration metrics with bootstrap standard deviations, and pooled coverage/set-size summaries at target 90% coverage.