CHAIN: CALIBRATED LLM FORECASTING VIA CAUSAL-TEMPORAL HYPERGRAPH INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models have achieved significant progress in event forecasting, yet their probability outputs exhibit systematic calibration bias that varies heterogeneously across different domains and question types, undermining the trustworthiness of probabilistic outputs for decision-making under uncertainty. However, existing calibration methods typically correct probability outputs after prediction is complete, without modeling the structural sources of bias within the prediction process itself. To address this challenge, we decompose probabilistic prediction over causal-temporal hypergraphs into three stages, evidence weighting, evidence aggregation, and source fusion, and propose **CHAIN**, which designs stage-specific mechanisms to mitigate bias at each stage: *(i)* modulating the temporal decay function by causal topological distance, *(ii)* aggregating approximately independent causal chains via Noisy-OR after direction-aware deduplication, and *(iii)* driving adaptive fusion by causal coverage and directional balance. Experimental results on cross-domain forecasting benchmarks show **CHAIN** outperforms existing methods in expected calibration error, Brier score, and accuracy. Our project is available[1].

1 INTRODUCTION

Large language models (LLMs) have achieved significant progress in event forecasting (Zhang et al., 2026; Paleka et al., 2025), advancing from direct prompting (Halawi et al., 2024; Yang et al., 2026; Wildman et al., 2025) and chain-of-thought (CoT) (Wei et al., 2022) reasoning to retrieval-augmented prediction (Wang et al., 2024a; Guan et al., 2024) and consistency-aware evaluation (Farquhar et al., 2024; Nakkiran et al., 2025), and further to structured reasoning over temporal knowledge graphs (Chen et al., 2024b), with prediction accuracy approaching human crowd levels (Baan et al., 2022; Peng et al., 2025). However, probability calibration remains a critical bottleneck for reliability, as model probability outputs exhibit persistent and widespread overconfidence (Xiao et al., 2025b), and calibration bias is heterogeneous across domains and question types (Niu et al., 2025; Luo et al., 2024), undermining the trustworthiness of probabilistic outputs for real-world decision-making under uncertainty (Zeng et al., 2025; Tao et al., 2025).

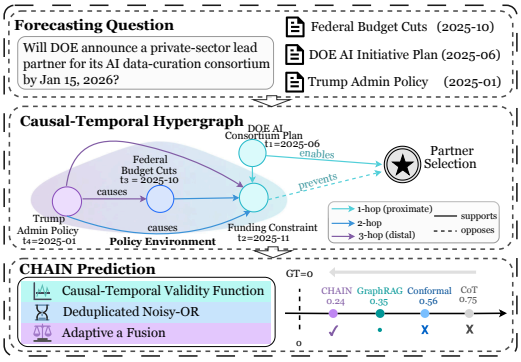


Figure 1: An illustration of the proposed **CHAIN**.

To calibrate predictions, post-hoc methods correct probability outputs after prediction (Geng et al., 2024), progressing from global parameter adjustment (Guo et al., 2017; Platt, 1999) to sample-adaptive (Xie et al., 2024a), semantic-level (Lamb et al., 2025), and alignment-aware (Xiao et al., 2025a) calibration. Non-parametric methods (Zadrozny & Elkan, 2002; 2001) and conformal prediction (Cherian et al., 2024; Manggala et al., 2024) provide distribution-free alternatives; prompt-based methods (Zhang et al., 2024) elicit verbalized confidence. These methods correct outputs post hoc (Geng et al., 2024) without modeling the structural sources of bias in the prediction.

[1] https://anonymous.4open.science/r/Chain-Ano-C654

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

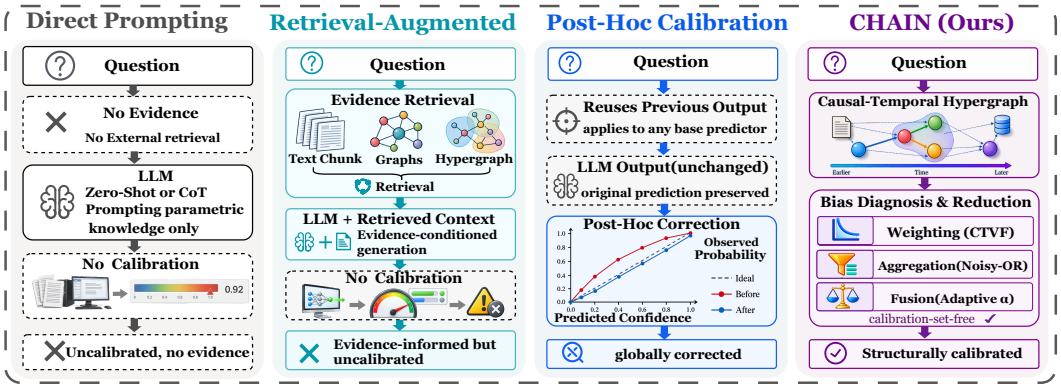


Figure 2: Inference-time approaches to calibrated event forecasting: existing methods and **CHAIN**.

However, these methods still face three key challenges in calibration: *(i)* Evidence at different causal distances is treated equally, with no distinction between **proximate causes** and **indirect background**, causing distal evidence to influence estimates as much as proximal ones. *(ii)* Multiple retrieval paths referencing the same fact participate in estimation without deduplication, causing the model to treat **redundant paths** as **independent evidence**, leading to systematically inflated confidence. *(iii)* The reliability of external evidence and internal estimates varies across questions, yet existing methods fuse them in a **fixed manner**, **lacking adaptive adjustment** according to evidence quality.

To address these challenges, we decompose probabilistic prediction over causal-temporal hypergraphs into three stages of evidence weighting, aggregation, and source fusion, and propose **CHAIN** (see Figure 1), which employs stage-specific mechanisms to mitigate bias at each stage. First, we propose the Causal-Temporal Validity Function (CTVF), modulating temporal decay by causal topological distance, capturing recency and causal relevance across retrieved evidence. Furthermore, we introduce direction-aware deduplication with Noisy-OR aggregation over approximately independent causal paths, reducing redundant confidence accumulation. In addition, we design adaptive fusion driven by causal coverage and directional balance, weighted by evidence quality to adjust source contributions.

We conduct experiments on 8 cross-domain forecasting benchmarks with multiple LLMs. Experimental results show **CHAIN** outperforms the baselines (see Figure 2) in both calibration and accuracy, with ablations showing that each removal degrades final calibration or prediction performance, validating the effectiveness of structural diagnosis over causal-temporal hypergraphs for probability calibration.

2 Related Work

Event Forecasting. Event forecasting assigns probabilities to events (Zou et al., 2022; Yuan et al., 2025). Direct prompting (Yuan et al., 2025; Lee et al., 2023), CoT (Wei et al., 2022), and retrieval methods (AutoCast++ (Yan et al., 2023); The Power of Simplicity (Zhang et al., 2025)) cover evidence-driven forecasting. Large-scale training (Lee et al., 2025), consistency checks (Paleka et al., 2024), and ensembles (Schoenegger et al., 2024) improve forecasts, while calibration varies across domains (Karkar & Chopra, 2025; Lu, 2025). GenTKG (Liao et al., 2024), INFER (Li et al., 2025), TPAR (Chen et al., 2024a), and zrLLM (Ding et al., 2024) perform graph extrapolation; context-aware modeling (Manggala et al., 2024) uses history. RAG (Gutiérrez et al., 2024; Luo et al., 2025) and multi-cognition aggregation (Wang et al., 2025) provide knowledge and multi-perspective fusion.

Probability Calibration. Calibration aligns predictions with observed frequencies (Geng et al., 2024). Post-hoc methods include temperature scaling (Guo et al., 2017), Platt scaling (Platt, 1999), ATS (Xie et al., 2024a), and internal consistency calibration (Xie et al., 2024b). Isotonic regression (Zadrozny & Elkan, 2002), histogram binning (Zadrozny & Elkan, 2001), and QA-calibration (Manggala et al., 2024) are non-parametric; ConU (Wang et al., 2024b) and enhanced conformal prediction (Cherian et al., 2024) provide distribution-free coverage guarantees. Just Ask (Tian et al., 2023) and Fidelity (Zhang et al., 2024) elicit confidence, while CFT (Xiao et al., 2025a) repairs alignment-induced degradation. Ca2KG (Ren et al., 2026) measures evidence-reliability variation, and Double-Calibration (Lu et al., 2026) calibrates evidence and reasoning confidence.

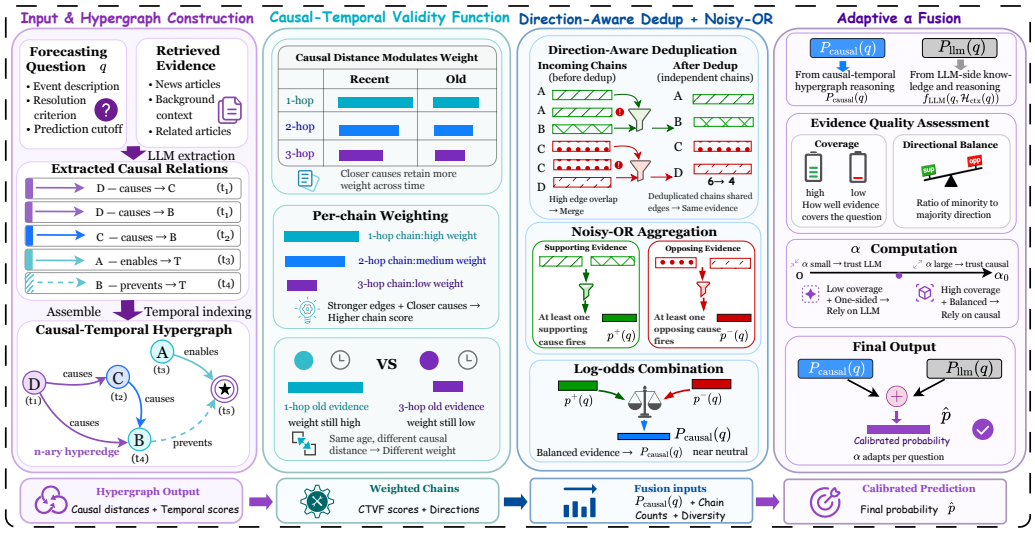


Figure 3: The overall framework of **CHAIN**, with CTVF, Noisy-OR aggregation, and adaptive fusion.

3 PRELIMINARIES

Event Forecasting and Calibration. At prediction cutoff t_c, the forecaster f outputs for question q:

$$p = f(q, t_c) \in [0, 1], \quad y \in \{0, 1\}. \tag{1}$$

Calibration quality is measured by the Expected Calibration Error (ECE), which is formally defined as $\text{ECE} = \sum_{b=1}^{B} (n_b/N)|\bar{p}_b - \bar{y}_b|$, where B is the number of equal-width bins, n_b, $\bar{p}_b$, $\bar{y}_b$ are the size, mean prediction, and empirical frequency of bin b respectively, and N is the total sample count.

Causal-Temporal Hypergraph. Evidence is organized as a causal-temporal hypergraph:

$$\mathcal{G} = (V, E_H, \mathcal{C}), \tag{2}$$

where V is the entity set; $(V_e, t, s) \in E_H$ stores $V_e \subseteq V$, the source-chunk timestamp t, and optional stored similarity $s \in [0, 1]$ when available; and $(u, v, \text{type}, w) \in \mathcal{C}$ is a directed edge, with $\text{type} \in \{\text{causes}, \text{enables}, \text{prevents}\}$ and $w \in [0, 1]$. For each chunk x, ϕ_{LLM} returns propositions and typed edges $\mathcal{C}_x$; the builder attaches t, s to form $\mathcal{P}_x$ before graph assembly:

$$E_H = \bigcup_x \mathcal{P}_x, \quad \mathcal{C} = \bigcup_x \mathcal{C}_x, \quad V = \bigcup_{(V_e, t, s) \in E_H} V_e. \tag{3}$$

At inference time, a cutoff-conditioned constructor derives an active graph $\mathcal{G}_q$ from the complete hypergraph, and both estimators use only this view:

$$\mathcal{H}_{\text{ctx}}(q) = \text{PromptContext}(\mathcal{G}_q, q, t_c), \qquad P_{\text{llm}} = f_{\text{LLM}}(q, \mathcal{H}_{\text{ctx}}(q)) \in [0, 1]. \tag{4}$$

Here $\mathcal{H}_{\text{ctx}}(q)$ contains narratives, trajectories, contradictions, causal-context chains, source events, and prior outcomes; outcome mapping projects the estimate onto the binary axis.

Causal Topological Distance. Using $\text{Enc}(\cdot)$, keyword-filtered dense reranking maps signed target descriptions from the reasoning-direction call to validated targets $\mathcal{T}(q) \subseteq V_q$, with $a_q(v) \in \{-1, +1\}$ marking the negative/positive axis side. For any node n in $\mathcal{G}_q$, define its shortest distance as:

$$d(n, q) = \min_{v \in \mathcal{T}(q)} d_{\mathcal{C}_q}(v, n), \tag{5}$$

where $d_{\mathcal{C}_q}$ is the unweighted Breadth-First Search (BFS) distance on the reverse of $\mathcal{C}_q$ (all types treated uniformly); a hyperedge e inherits $d(e, q) = \min_{v \in V_e} d(v, q)$; unreachable nodes have $d = \infty$.

4 METHOD: CHAIN

In this section, we introduce **CHAIN** (see Figure 3), which addresses calibration bias at three pipeline stages through the Causal-Temporal Validity Function for evidence weighting, direction-aware Noisy-OR for chain aggregation, and causal-coverage-balanced adaptive fusion for source combination.

4.1 CAUSAL-TEMPORAL VALIDITY FUNCTION

The CTVF assigns each piece of evidence in $\mathcal{G}_q$ a validity $V_{\mathrm{CTVF}} \in (0, 1]$ reflecting temporal recency and causal proximity to $\mathcal{T}(q)$, propagated along causal paths into chain-level confidences $\Phi(\pi, q)$.

Causal-Proximity Exponent. Using the causal topological distance $d(n, q)$ defined in the preliminaries, a proximity exponent $\rho(n, q) \in [0, 1]$ is obtained for every node n in the active graph with $n \in V_q \cup E_{H,q}$ by normalizing $d(n, q)$ and clipping it against the reachability horizon $d_{\max} \in \mathbb{Z}^+$:

$$\rho(n, q) = \min\left(\frac{d(n, q)}{d_{\max}},\ 1\right). \tag{6}$$

Thus $\rho = 0$ at the target, $\rho \in (0, 1)$ within the horizon, and $\rho = 1$ at or beyond it or for unreachable nodes; larger ρ induces faster temporal decay for distant evidence.

Temporal Recency Component. For a hyperedge $e = (V_e, t, s) \in E_{H,q}$ with a parseable timestamp $t < t_c$, let $\Delta t(e) = \max\{1, (t_c - t)_{\mathrm{days}}\}$. The temporal recency component $V_{\mathrm{rec}}(e, q) \in (0, 1]$ is modeled as an exponential function of the logarithmic span, whose decay rate is scaled by $\rho(e, q)$:

$$V_{\mathrm{rec}}(e, q) = \exp\Big(-\tfrac{1}{2}\,\rho(e, q) \cdot \ln \Delta t(e)\Big). \tag{7}$$

At $\rho(e, q) = 0$ the exponent vanishes and $V_{\mathrm{rec}} = 1$, so causally direct evidence retains full recency; at $\rho(e, q) = 1$, $V_{\mathrm{rec}} = 1/\sqrt{\Delta t(e)}$, yielding maximal temporal decay for distant evidence.

Recurrence-Salience Component. Let $N(e)$ count distinct admitted record IDs for e's canonicalized proposition key, and let $M_r(e) = s \in [0, 1]$ be an optional stored similarity. The centered logistic coefficient $\mu(N)$ amplifies co-supported evidence and suppresses low-signal singletons:

$$\mu(N) = \frac{1}{1 + \exp(-\sqrt{N}/\eta)} - \tfrac{1}{2}, \qquad V_{\mathrm{sal}}(e) = \mu(N(e)) \cdot \sqrt{N(e)} \cdot M_r(e), \tag{8}$$

where $\eta > 0$ controls saturation; $\sqrt{N(e)}$ yields sub-linear recurrence amplification, and $M_r(e)$ optionally down-weights weak semantic matches under weak similarity.

CTVF Validity. For valid pre-cutoff timestamps, the two components are summed and capped at 1:

$$V_{\mathrm{CTVF}}(e, q) = \min\big(V_{\mathrm{rec}}(e, q) + V_{\mathrm{sal}}(e),\ 1\big). \tag{9}$$

Known $t \geq t_c$ is excluded; missing or unparseable t uses v_{fb} only with a verified pre-cutoff availability bound. Entity scores average admitted incident-hyperedge values, with fallback 0.5 for aggregation.

Edge-Level Weight. For each directed causal edge $c = (u, v, \mathrm{type}, w) \in \mathcal{C}_q$, the CTVF-weighted edge scalar $\omega(c) \in [0, 1]$ is the mean of the endpoint validities raised to an edge-level exponent ρ_c:

$$\omega(c) = \big(\tfrac{1}{2}\big(V_{\mathrm{CTVF}}(u, q) + V_{\mathrm{CTVF}}(v, q)\big)\big)^{\rho_c}, \qquad \rho_c = \min\big(\tfrac{d(u,q)+d(v,q)}{2\,d_{\max}},\ 1\big) \in [0, 1]. \tag{10}$$

Edges located near the target receive a small ρ_c and keep their endpoint mean close to 1; meanwhile, distal edges have $\rho_c \to 1$ and inherit the uncorrected raw mean validity $\omega(c)$ of their endpoints.

Target-Termination Gate. Writing $c_\ell = (v_{\ell-1}, v_\ell, \mathrm{type}_\ell, w_\ell)$, a node-simple directed chain $\pi = (c_1, \ldots, c_L)$, $1 \leq L \leq d_{\max}$, is retained only when it terminates at an oriented target:

$$\iota_q(\pi) = \mathbb{1}[v_L \in \mathcal{T}(q)]\,. \tag{11}$$

Chain-Level Confidence. For every score-qualified candidate chain, the confidence $\Phi(\pi, q) \in [0, 1]$ is the target-gated form of the ungated causal-chain score $\widetilde{\Phi}(\pi, q)$:

$$\widetilde{\Phi}(\pi, q) = \prod_{\ell=1}^{L} w_\ell \omega(c_\ell), \qquad \Phi(\pi, q) = \iota_q(\pi)\widetilde{\Phi}(\pi, q). \tag{12}$$

Forward breadth-first search expands at most $F_{\max}$ outgoing edges per node and prunes prefixes whose ungated score is below τ_Φ. Only score-qualified chains terminating at oriented targets enter the top-B_π pool Π_q; otherwise the branch abstains from contributing evidence to the final forecast.

Proposition 1. *Under normalized weighting, CTVF lowers weighting error versus uniform weighting by favoring lower-error proximal evidence without increasing the weighted error within either subset.*

Proof. We provide experimental results in Section 5.4 and theoretical proofs in Appendix A.1. □

4.2 Evidence Aggregation via Direction-aware Noisy-OR

Direction-Aware Noisy-OR reduces redundant paths from acting as independent evidence through polarity-wise deduplication and neutral-anchored log-odds aggregation of directional evidence.

Causal Polarity Encoding. Each retained chain $\pi = (c_1, \ldots, c_L)$ receives a binary question-conditioned polarity $\delta_q(\pi) \in \{+1, -1\}$. Let $P(\pi)$ count its prevents edges; target orientation and relation parity determine the polarity partitions:

$$\delta_q(\pi) = a_q(v_L)(-1)^{P(\pi)}, \qquad \Pi_q^+ = \{\pi \in \Pi_q : \delta_q(\pi) = +1\}, \quad \Pi_q^- = \{\pi \in \Pi_q : \delta_q(\pi) = -1\}. \tag{13}$$

The endpoint orientation fixes the target's positive/negative axis side and final polarity, while each $\mathrm{prevents}$ edge reverses the direction propagated along the chain during aggregation and fusion.

Same-Polarity Deduplication. Two same-polarity chains π_i, π_j are treated as non-independent when their edge sets overlap excessively. Let $E(\pi)$ denote the edge set of chain π; the Jaccard overlap is:

$$J(\pi_i, \pi_j) = \frac{|E(\pi_i) \cap E(\pi_j)|}{|E(\pi_i) \cup E(\pi_j)|}. \tag{14}$$

Within each polarity set $\Pi_q^{\pm}$, the chains are sorted in descending order of $\Phi(\pi, q)$ and greedily added to $\widehat{\Pi}_q^{\pm}$ for subsequent greedy selection: a chain π is admitted iff

$$\max_{\pi' \in \widehat{\Pi}_q^{\pm}} J(\pi, \pi') \le \theta, \tag{15}$$

Under the convention $\max \varnothing := 0$ and overlap threshold $\theta \in (0, 1)$, the procedure greedily retains each chain whose overlap with all previously retained chains is at most θ, and discards the rest.

Polarity-Wise Noisy-OR. The surviving chains in each channel are combined under a within-polarity approximate-independence assumption. The resulting channel scores are:

$$p^+(q) = 1 - \prod_{\pi \in \widehat{\Pi}_q^+} \big(1 - \Phi(\pi, q)\big), \qquad p^-(q) = 1 - \prod_{\pi \in \widehat{\Pi}_q^-} \big(1 - \Phi(\pi, q)\big), \tag{16}$$

Where $p^+(q), p^-(q) \in [0, 1]$ are the positive- and negative-channel aggregates, they summarize the surviving evidence in the two directional channels for downstream fusion.

Neutral-Anchored Log-Odds. For fixed $\varepsilon = 10^{-6}$, define $g_\varepsilon(p) = \mathrm{logit}(\min(1 - \varepsilon, \max(\varepsilon, (1 + p)/2)))$ to map zero directional evidence to neutral log-odds and obtain the causal probability:

$$P_{\mathrm{causal}}(q) = \sigma\Big(g_\varepsilon(p^+(q)) - g_\varepsilon(p^-(q))\Big), \tag{17}$$

where $\mathrm{logit}(p) = \ln(p/(1-p))$ and $\sigma(z) = 1/(1 + e^{-z})$. The rule returns $1/2$ for empty or equally strong channels, increases continuously with p^+, and decreases continuously with p^-, and remains neutral when directional evidence is absent or balanced across both channels.

Proposition 2. *Direction-aware deduplication can eliminate the same-polarity Noisy-OR probability inflation induced by overlap-identified redundant causal chains.*

Proof. We provide experimental results in Section 5.4 and theoretical proofs in Appendix A.2. □

4.3 Source Fusion via Causal-Coverage-Balanced Adaptive Fusion

Because evidence reliability varies, fixed weighting of P_{llm} and P_{causal} may be suboptimal; adaptive fusion derives bounded $\alpha(q)$ from retained evidence and blends them into $\hat{p}$.

Evidence Reliability. Let $k = |\widehat{\Pi}_q^+| + |\widehat{\Pi}_q^-|$ denote the number of approximately independent chains after deduplication. A logarithmic reliability score $\kappa(k) \in [0, 1]$ saturates as more such chains accumulate:

$$\kappa(k) = \min\left(\frac{\ln(1+k)}{\ln(1+k_{\mathrm{sat}})},\ 1\right), \tag{18}$$

where $k_{\mathrm{sat}} \in \mathbb{Z}^+$ is a saturation anchor controlling the logarithmic curve. Hence $\kappa(k)$ grows quickly with the first few independent chains and gradually saturates once their count k exceeds k_{sat}.

Directional Balance. A polarity-imbalance factor $\beta(q) \in [0,1]$ explicitly quantifies the relative evenness of the positive- and negative-polarity chain sets $\widehat{\Pi}_q^+$ and $\widehat{\Pi}_q^-$ via:

$$\beta(q) = \frac{\min\big(|\widehat{\Pi}_q^+|,\, |\widehat{\Pi}_q^-|\big)}{\max\big(|\widehat{\Pi}_q^+|,\, |\widehat{\Pi}_q^-|,\, 1\big)}. \tag{19}$$

A value of $\beta(q) = 0$ indicates fully one-sided evidence, whereas $\beta(q) = 1$ indicates a perfectly balanced split between sides. To prevent one-sided evidence from being entirely discarded, $\beta(q)$ is mapped through an affine floor $\widetilde{\beta}(q) = \beta_0 + (1-\beta_0)\,\beta(q) \in [\beta_0, 1]$ with floor parameter $\beta_0 \in (0,1)$.

Causal Coverage. The causal coverage $\Omega(q) \in [0,1]$ aggregates the count k, the mean confidence $\bar{\Phi}(q) = \frac{1}{k}\sum_{\pi \in \widehat{\Pi}_q^+ \cup \widehat{\Pi}_q^-} \Phi(\pi, q)$ (0 when $k = 0$), the reliability $\kappa(k)$, and the floored balance $\widetilde{\beta}(q)$:

$$\Omega(q) = \min\Big(\frac{k \cdot \bar{\Phi}(q) \cdot \kappa(k) \cdot \widetilde{\beta}(q)}{Z},\ 1\Big), \tag{20}$$

where $Z > 0$ is a predefined normalization constant. High $\Omega(q)$ corresponds to abundant, strong, and balanced causal evidence, whereas low $\Omega(q)$ corresponds to sparse, weak, or one-sided evidence.

Adaptive Fusion Weight. A coverage-driven sigmoid maps $\Omega(q)$ to $\alpha(q) \in [0, \alpha_0]$:

$$\alpha(q) = \mathbb{1}[k > 0] \cdot \min\big(2\alpha_b\, \sigma\big(\zeta\,(\Omega(q) - \Omega_0)\big),\ \alpha_0\big), \tag{21}$$

where $\mathbb{1}[\cdot]$ is an indicator and σ is the sigmoid; $\alpha_b > 0$ is the mapping scale, $0 < \alpha_0 \leq 1$ is the fusion upper bound, Ω_0 is the coverage activation threshold, and $\zeta > 0$ is the sigmoid-transition sharpness.

Final Forecast. The final probability $\hat{p} \in [0,1]$ is the convex combination weighted by $\alpha(q)$:

$$\hat{p} = \alpha(q) \cdot P_{\text{causal}}(q) + \big(1 - \alpha(q)\big) \cdot P_{\text{llm}}(q). \tag{22}$$

The causal contribution is zero when no chain is retained and otherwise increases with causal coverage up to α_0, while the complementary LLM contribution decreases accordingly.

Proposition 3. *Adaptive fusion using causal coverage and directional balance down-weights the causal estimate when evidence is unreliable, and can yield lower squared error than fixed weighting.*

Proof. We provide experimental results in Section 5.6 and theoretical proofs in Appendix A.3. □

5 EXPERIMENTS

In this section, we answer these following research questions (RQ): **RQ1:** whether **CHAIN** outperforms the baselines? **RQ2:** its generalization to out-of-distribution (OOD) datasets? **RQ3:** its components' contributions to overall performance? **RQ4:** whether structural calibration outperforms post-hoc methods? **RQ5:** whether the adaptive fusion weight α outperforms any fixed alternative?

5.1 EXPERIMENTAL SETUP

Datasets. We use eight datasets: **AI-Futures** (LightningRodLabs, 2025a), **Metaculus** (Chandak et al., 2026), **Polymarket** (Polymarket, 2024), **Future-as-Label (FaL)** (Turtel et al., 2026), **Clinical Trial Outcomes (CTO)** (Gao et al., 2024), **FOReCAst** (Yuan et al., 2025), **Golf-Forecasting** (LightningRodLabs, 2025b), and **KalshiBench** (Nel, 2025). More details are shown in Appendix D.

Baselines. We compare **CHAIN** with 13 baselines: **LLM Direct**, **LLM CoT**, **AutoCast**, **NaiveRAG**, **GraphRAG**, **LightRAG**, **HippoRAG**, **HyperGraphRAG**, **TempScaling**, **PlattScaling**, **IsotonicRegression**, **Histogram Binning**, and **ConformalAdj**. More details are illustrated in Appendix E.

Evaluation Metrics. We use seven evaluation metrics, including **Expected Calibration Error (ECE)** (Guo et al., 2017), **Adaptive Calibration Error (ACE)** (Nixon et al., 2019), **Maximum Calibration Error (MCE)** (Guo et al., 2017), **Reliability (Rel)** (Murphy, 1973), **Negative Log-Likelihood (NLL)**, **Brier Score** (Brier, 1950), and **Accuracy (Acc)**. More details see Appendix F.

Implementation. We construct the causal-temporal hypergraph via **GPT-4o-mini** (Hurst et al., 2024) and evaluate **CHAIN** three LLMs, including **GPT-4o-mini**, **Gemini-2.0-Flash** (Kavukcuoglu, 2025), and **DeepSeek-V3** (DeepSeek-AI et al., 2025). More details are illustrated in Appendix G.

Table 1: Main results of **CHAIN** and the baselines across four in-distribution forecasting datasets, which are reported on three LLMs with **best** in bold. All values are in %.

Method	AI-Futures			Metaculus			Polymarket			Future-as-Label			Average		
	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑
GPT-4o-mini															
LLM Direct	15.00	25.69	59.38	22.27	28.79	52.34	20.04	23.19	70.31	21.29	29.97	53.91	19.65	26.91	58.98
LLM CoT	12.89	20.27	71.09	14.22	23.58	64.06	17.42	24.88	64.06	22.62	29.74	51.56	16.79	24.62	62.70
AutoCast	32.85	32.49	43.75	37.77	37.40	42.19	22.38	25.56	68.75	34.92	35.30	49.22	31.98	32.69	50.98
NaiveRAG	26.99	28.79	43.75	28.71	30.72	46.88	14.18	23.96	68.75	36.64	36.62	40.62	26.63	30.02	50.00
GraphRAG	33.05	34.39	39.84	25.35	28.81	52.34	18.67	23.74	68.75	33.83	36.26	41.41	27.72	30.80	50.59
LightRAG	36.29	36.38	41.41	28.44	29.67	51.56	25.94	25.24	67.19	21.68	27.19	60.16	28.09	29.62	55.08
HippoRAG	20.72	24.80	59.38	28.59	32.64	49.22	16.45	23.22	70.31	29.69	32.77	53.12	23.86	28.36	58.01
HyperGraphRAG	32.70	34.71	45.31	31.64	32.46	46.09	22.70	25.69	67.19	25.55	28.84	57.03	28.14	30.43	53.91
TempScaling	14.14	23.84	59.38	24.00	29.69	52.34	18.15	22.80	70.31	16.99	27.57	53.91	18.32	25.98	58.98
PlattScaling	31.84	31.77	33.59	24.98	29.67	52.34	18.24	23.12	70.31	27.37	30.99	42.19	25.61	28.89	49.61
IsotonicReg	33.06	33.11	39.06	22.45	29.40	52.34	18.33	24.91	70.31	33.41	36.40	42.19	26.81	30.96	50.98
HistogramBin	32.67	34.60	39.06	26.31	30.84	50.78	17.23	22.93	71.09	33.68	37.83	43.75	27.47	31.55	51.17
ConformalAdj	15.41	24.08	59.38	15.04	24.81	52.34	14.83	23.58	70.31	14.72	24.97	53.91	15.00	24.36	58.98
CHAIN (ours)	**9.45**	**17.15**	**77.34**	**7.52**	**18.77**	**71.88**	**13.74**	**19.54**	**71.88**	**10.20**	**19.68**	**67.97**	**10.23**	**18.79**	**72.27**
Gemini-2.0-Flash															
LLM Direct	9.18	20.32	67.97	21.89	25.57	58.59	15.39	22.57	71.88	28.52	33.23	49.22	18.74	25.42	61.91
LLM CoT	9.18	20.41	67.19	13.43	21.98	65.62	15.00	20.46	71.88	19.77	28.34	53.91	14.35	22.80	64.65
AutoCast	13.63	23.30	60.94	13.91	24.78	58.59	21.48	25.10	35.94	15.12	26.80	57.03	16.04	25.00	53.12
NaiveRAG	18.44	25.04	42.97	13.91	22.76	63.28	19.69	25.35	32.03	14.70	25.28	57.81	16.68	24.61	49.02
GraphRAG	12.15	22.01	57.81	14.97	23.56	59.38	23.19	24.27	68.75	19.27	29.88	45.31	17.39	24.93	57.81
LightRAG	14.96	23.03	57.81	14.49	21.48	66.41	29.34	26.86	46.09	15.82	24.76	58.59	18.65	24.03	57.23
HippoRAG	10.39	19.49	67.19	16.02	23.94	60.94	26.80	26.00	39.84	17.30	25.51	59.38	17.63	23.73	56.84
HyperGraphRAG	13.40	24.31	56.25	14.45	22.51	64.84	23.75	24.79	35.94	17.15	23.62	56.25	17.19	23.81	53.32
TempScaling	8.13	20.21	67.97	15.62	23.28	58.59	16.55	22.78	71.88	18.67	28.68	49.22	14.74	23.74	61.91
PlattScaling	26.96	27.19	43.75	20.71	26.29	57.81	14.90	21.80	71.88	28.13	33.68	44.53	22.67	27.24	54.49
IsotonicReg	26.27	27.48	54.69	22.80	28.37	64.06	23.43	25.48	43.75	29.47	35.52	44.53	25.49	29.21	51.76
HistogramBin	25.82	28.70	54.69	24.67	34.65	53.91	26.49	29.39	44.53	29.39	35.89	42.19	26.59	32.16	48.83
ConformalAdj	13.29	22.44	67.97	13.20	23.61	58.59	17.11	23.31	71.88	11.74	25.53	49.22	13.83	23.72	61.91
CHAIN (ours)	**7.10**	**17.19**	**73.44**	**7.05**	**17.73**	**74.22**	**14.87**	**18.96**	**75.00**	**9.72**	**19.96**	**68.75**	**9.68**	**18.46**	**72.85**
DeepSeek-V3															
LLM Direct	9.34	21.17	68.75	20.90	26.69	54.69	15.43	20.71	74.22	19.69	29.16	54.69	16.34	24.43	63.09
LLM CoT	11.48	21.62	66.41	12.49	18.86	71.88	14.70	20.63	71.88	15.72	26.99	59.38	13.60	22.03	67.38
AutoCast	15.43	24.37	53.91	19.73	27.06	56.25	20.43	24.73	43.75	17.58	28.47	51.56	18.29	26.16	51.37
NaiveRAG	11.91	21.31	63.28	19.88	25.19	57.81	16.95	23.54	58.59	24.34	29.42	47.66	18.27	24.87	56.84
GraphRAG	16.99	23.57	58.59	18.98	25.64	58.59	15.12	20.44	73.44	24.34	30.76	47.66	18.86	25.10	59.57
LightRAG	17.97	24.14	58.59	17.03	22.89	64.06	15.90	23.28	66.41	19.61	26.66	57.03	17.63	24.24	61.52
HippoRAG	10.39	19.24	69.53	16.76	23.86	64.06	15.66	20.86	71.09	11.76	20.88	67.97	13.64	21.21	68.16
HyperGraphRAG	19.73	25.90	55.47	23.98	25.02	59.38	16.09	20.78	71.09	12.77	19.43	71.09	18.14	22.78	64.26
TempScaling	8.85	20.79	68.75	16.05	24.44	54.69	16.42	21.58	74.22	13.41	26.23	54.69	13.68	23.26	63.09
PlattScaling	25.71	26.76	64.84	24.71	28.75	53.12	13.82	20.86	75.00	28.54	31.79	44.53	23.20	27.04	59.38
IsotonicReg	24.72	25.57	68.75	25.48	30.33	54.69	27.55	26.89	33.59	31.05	34.61	44.53	27.20	29.35	50.39
HistogramBin	30.53	27.63	53.12	26.62	30.51	53.12	29.59	28.42	32.81	32.65	35.49	50.78	29.85	30.51	47.46
ConformalAdj	13.53	22.80	68.75	13.27	24.12	54.69	18.96	23.01	74.22	11.72	24.81	54.69	14.37	23.68	63.09
CHAIN (ours)	**8.54**	**17.34**	**76.56**	**9.72**	**18.05**	**73.44**	**13.74**	**17.41**	**79.69**	**10.09**	**18.82**	**72.66**	**10.52**	**17.91**	**75.59**

5.2 MAIN RESULTS (RQ1)

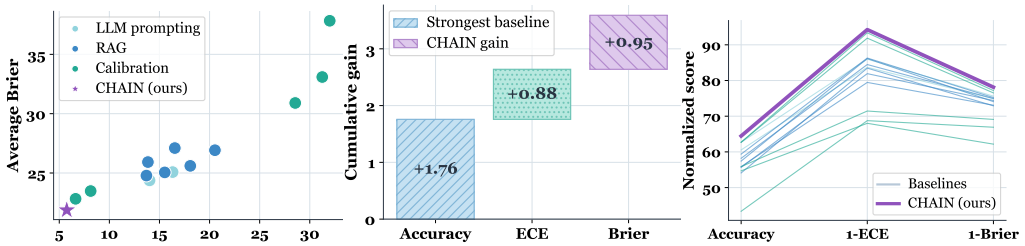
Table 1 shows the main results for three LLMs across four in-distribution datasets. *(i)* **Joint calibration and accuracy. CHAIN** has the lowest ECE and Brier and the highest accuracy, improving calibration without sacrificing performance. *(ii)* **Prompting and retrieval remain insufficient.** Direct prompting has higher calibration error, while retrieval gains vary, so added evidence alone does not ensure calibrated forecasts. *(iii)* **Post-hoc correction leaves a gap.** Post-hoc methods retain higher calibration error and sometimes reduce accuracy, showing that output correction alone is insufficient. *(iv)* **Backbone-agnostic gains. CHAIN** leads across all three backbones, with lower average ECE and Brier and higher accuracy, showing consistent gains across these evaluated models.

5.3 OUT-OF-DISTRIBUTION GENERALIZATION (RQ2)

Table 2 and Figure 4 show the results for the same three LLMs across four out-of-distribution datasets. *(i)* **Calibration transfers to OOD. CHAIN** attains the lowest ECE and Brier, maintaining its calibration advantage as forecasting tasks move beyond the in-distribution benchmarks. *(ii)* **Accuracy remains highest under shift. CHAIN** leads in the OOD results, showing that lower calibration error

Table 2: Performance comparison of **CHAIN** and the baselines on 4 OOD datasets across 2 LLMs.

Method	CTO			FOReCAst			Golf-Forecasting			KalshiBench			Average		
	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑
GPT-4o-mini															
LLM Direct	18.83	27.61	55.47	12.27	23.43	67.19	19.77	27.18	53.91	20.04	27.81	59.38	17.72	26.51	58.98
LLM CoT	18.12	26.88	56.25	16.37	24.78	60.16	11.25	25.13	60.16	16.51	26.71	57.81	15.56	25.88	58.59
AutoCast	42.50	43.07	42.19	26.09	28.93	59.38	45.35	43.72	47.66	44.30	44.26	41.41	39.56	39.99	47.66
NaiveRAG	30.04	32.62	42.97	30.35	33.02	50.00	24.14	30.31	48.44	30.94	34.34	44.53	28.87	32.57	46.48
GraphRAG	23.79	29.67	48.44	22.27	28.45	57.81	23.16	30.17	48.44	25.70	31.29	50.00	23.73	29.90	51.17
LightRAG	31.60	33.38	42.97	13.79	23.92	68.75	30.00	34.43	45.31	27.62	29.51	55.47	25.75	30.31	53.12
HippoRAG	27.27	30.53	44.53	18.63	25.82	64.06	31.64	34.07	46.09	33.20	34.35	50.78	27.69	31.19	51.37
HyperGraphRAG	33.09	34.63	42.19	21.41	28.23	59.38	31.41	34.32	45.31	27.54	31.56	52.34	28.36	32.19	49.80
TempScaling	9.57	24.86	55.47	10.52	22.38	67.19	13.62	24.82	53.91	13.45	24.76	59.38	11.79	24.21	58.98
PlattScaling	31.70	34.35	42.19	31.50	32.99	46.88	34.90	36.12	44.53	37.15	36.89	39.84	33.81	35.09	43.36
IsotonicReg	29.93	33.28	42.19	34.82	34.56	46.88	32.32	34.55	44.53	37.69	38.53	39.84	33.69	35.23	43.36
HistogramBin	34.10	35.92	46.09	31.42	31.50	58.59	31.24	33.71	50.78	33.09	36.25	50.78	32.46	34.34	51.56
ConformalAdj	4.45	24.53	55.47	14.43	23.06	67.19	8.79	24.59	53.91	10.73	24.19	59.38	9.60	24.09	58.98
CHAIN (ours)	**4.36**	**22.63**	**64.84**	**9.95**	**19.83**	**69.53**	**8.19**	**23.64**	**62.50**	**9.85**	**23.39**	**61.72**	**8.09**	**22.37**	**64.65**
Gemini-2.0-Flash															
LLM Direct	12.66	23.16	63.28	9.38	20.00	71.88	13.67	26.92	57.81	21.28	26.87	60.16	14.25	24.23	63.28
LLM CoT	11.88	23.95	63.28	11.05	21.98	67.19	23.01	30.51	50.00	19.98	25.99	60.16	16.48	25.61	60.16
AutoCast	7.93	24.92	57.81	11.76	24.19	56.25	11.48	26.19	58.59	11.37	26.36	56.25	10.63	25.41	57.23
NaiveRAG	9.62	24.56	55.47	20.10	26.82	43.75	5.82	24.05	55.47	11.73	25.29	58.59	11.82	25.18	53.32
GraphRAG	7.70	24.44	54.69	9.88	21.63	69.53	7.30	24.59	60.16	18.24	26.26	53.12	10.78	24.23	59.38
LightRAG	11.87	26.09	50.00	17.54	22.76	68.75	6.84	25.14	53.12	13.44	23.35	56.25	12.42	24.34	57.03
HippoRAG	12.89	24.30	55.47	16.60	21.09	70.31	12.28	28.16	43.75	21.25	25.65	50.78	15.76	24.80	55.08
HyperGraphRAG	7.07	23.62	59.38	17.30	24.49	61.72	7.73	24.45	58.59	12.03	24.20	56.25	11.04	24.19	58.98
TempScaling	4.71	22.27	63.28	10.57	19.74	71.88	7.40	24.84	57.81	14.47	23.56	60.16	9.29	22.60	63.28
PlattScaling	29.91	31.62	42.19	27.91	28.22	46.88	31.53	34.05	44.53	31.83	33.08	39.84	30.30	31.74	43.36
IsotonicReg	29.65	32.14	44.53	23.38	27.25	56.25	31.80	34.88	45.31	33.60	34.62	48.44	29.61	32.22	48.63
HistogramBin	45.71	42.75	41.41	39.90	37.49	43.75	35.27	38.80	50.78	39.51	41.14	47.66	40.10	40.04	45.90
ConformalAdj	7.25	23.33	63.28	15.49	21.95	71.88	5.63	24.45	57.81	9.70	23.59	60.16	9.51	23.33	63.28
CHAIN (ours)	**4.54**	**22.15**	**64.84**	**8.05**	**18.95**	**74.22**	**5.41**	**22.42**	**63.28**	**9.63**	**23.17**	**65.62**	**6.91**	**21.67**	**66.99**



(a) Pareto calibration trade-off (b) Cumulative gains (c) Cross-metric performance

Figure 4: Calibration and predictive performance of **CHAIN** for four OOD datasets on DeepSeek-V3.

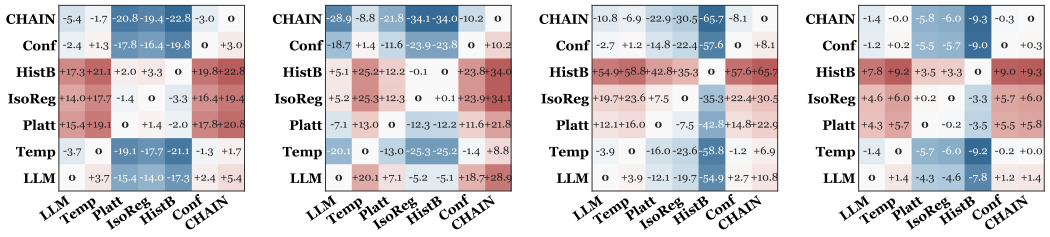
does not come at the expense of prediction accuracy under distribution shift. *(iii)* **Gains persist across backbones.** Calibration and accuracy advantages recur across models, indicating that the improvements are not tied to one model's forecasting capability or one source. *(iv)* **Joint calibration–accuracy gains.** Lower average ECE and Brier accompany higher accuracy, indicating balanced improvements in both the quality of probability estimates and the correctness of event predictions.

5.4 Ablation Study (RQ3)

As shown in Table 3, **CHAIN**'s three components each contribute to calibration and prediction accuracy: *(i)* **CTVF and Noisy-OR improve calibration.** Removing either component raises ECE on every dataset, showing that temporal-causal validity weighting and direction-aware aggregation both improve the quality of predicted probabilities. *(ii)* **Noisy-OR safeguards OOD accuracy.** Removing Noisy-OR causes the largest accuracy drop on each OOD dataset, demonstrating the role of direction-aware aggregation in maintaining OOD prediction accuracy. *(iii)* **Adaptive α tightens residual calibration.** Replacing per-instance fusion with a fixed weight raises ECE on every dataset, showing that question-level reweighting improves calibration and predictive accuracy.

Table 3: Ablation study of **CHAIN**, where *w/o* means without the corresponding component.

Method	AI-Futures			Metaculus			Polymarket			Future-as-Label		
	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑
w/o CTVF	11.68	19.43	73.44	10.65	19.59	72.66	19.73	17.54	75.00	10.91	23.15	63.28
w/o Noisy-OR	12.79	19.37	75.78	12.61	19.72	71.09	16.22	17.86	78.91	11.03	19.78	71.88
w/o Adaptive α	11.25	17.66	75.78	10.40	18.27	72.66	14.39	17.66	77.34	10.16	19.68	71.88
CHAIN (Full)	**8.54**	**17.34**	**76.56**	**9.72**	**18.05**	**73.44**	**13.74**	**17.41**	**79.69**	**10.09**	**18.82**	**72.66**
Method	**CTO**			**FOReCAst**			**Golf-Forecasting**			**KalshiBench**		
	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑
w/o CTVF	7.24	22.92	60.94	8.73	17.95	73.44	7.44	24.59	57.81	13.33	24.42	60.94
w/o Noisy-OR	5.02	23.47	60.16	8.43	18.87	72.66	7.76	24.02	57.03	9.13	23.20	60.16
w/o Adaptive α	3.72	22.85	61.72	7.35	18.17	73.44	6.52	24.00	57.81	9.04	22.98	61.72
CHAIN (Full)	**3.61**	**22.73**	**62.50**	**6.98**	**17.90**	**74.22**	**5.15**	**23.91**	**58.59**	**7.25**	**22.95**	**62.50**

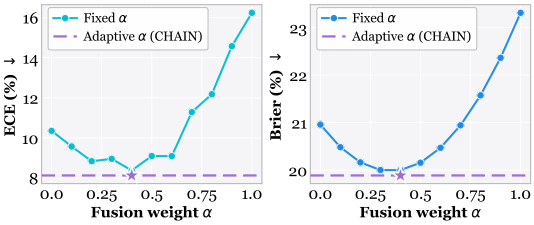


(a) ACE. (b) MCE. (c) NLL ($\times 100$). (d) Rel.

Figure 5: Comparison of **CHAIN** and post-hoc calibration methods on DeepSeek-V3 (lower is **better**).

5.5 Structural Calibration vs Post-Hoc Fitting (RQ4)

Figure 5 shows that, on the pooled eight-dataset evaluation, **CHAIN** outperforms all post-hoc baselines across all four metrics. *(i)* **Pairwise dominance.** **CHAIN** attains the lowest ACE, MCE, NLL, and Rel. *(ii)* **Worst-bin calibration error suppressed.** **CHAIN** reduces MCE relative to every post-hoc baseline, combining lower worst-bin error with improved average probability quality. *(iii)* **Post-hoc fitting can exacerbate miscalibration.** Platt, Isotonic, and Histogram Binning have higher ACE and Rel than LLM Direct, while **CHAIN** records the lowest values for both calibration metrics.



(a) ECE. (b) Brier.

Figure 6: DeepSeek-V3 macro-average ECE/Brier over eight ID/OOD datasets: adaptive vs. fixed α.

5.6 Adaptive α vs Fixed (RQ5)

As shown in Figure 6, the fixed-α sensitivity analysis reveals three patterns in the eight-dataset macro-average: *(i)* **Point-estimate comparison.** Adaptive α yields lower ECE and Brier than every evaluated fixed α, revealing a consistent advantage over fixed fusion. *(ii)* **Fixed-weight sensitivity.** ECE and Brier vary across fixed weights, with excessively large fixed α degrading both metrics, revealing sensitivity to fusion choice. *(iii)* **Adaptive fusion.** These results support question-level weighting over globally fixed fusion when evidence quality differs across questions and datasets.

6 Conclusion

In this work, we propose **CHAIN**, a causal-temporal hypergraph framework for diagnosing and mitigating LLM miscalibration in event forecasting. Experimental results on eight datasets across multiple LLM backbones illustrate that **CHAIN** consistently improves calibration and prediction accuracy over the evaluated baselines, validating structural calibration within the prediction process.

AI USE STATEMENT

In this work, we used generative AI tools to reformat the experimental dataset, but not to clean the data. Apart from this task, we did not use generative AI for any other required research tasks. Generating synthetic datasets, translation, and qualitative or thematic analysis were not applicable to this work. We also used GPT-6 Astra to improve the manuscript's language, grammar, and readability and OpenAI's ChatGPT Images image-generation feature to assist with some figures. We reviewed and revised all AI-assisted outputs as necessary. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public benchmark datasets without new human-subject recruitment or personal data collection. Forecasts are for research evaluation and not high-impact decisions.

REPRODUCIBILITY STATEMENT

The paper and appendix provide methods, prompts, proofs, datasets, evaluation settings, and extended results. Code and supporting materials are provided for reproduction.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1892–1915, 2022.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future, 2026. URL https://arxiv.org/abs/2512.25070.

Kai Chen, Ye Wang, Yitong Li, Aiping Li, Han Yu, and Xin Song. A unified temporal knowledge graph reasoning model towards interpolation and extrapolation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 117–132, 2024a.

Kai Chen, Ye Wang, Xin Song, Siwei Chen, Han Yu, and Aiping Li. Temporal knowledge graph extrapolation via causal subhistory identification. In *IJCAI*, pp. 3298–3306, 2024b.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu,

Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Zifeng Ding, Heling Cai, Jingpei Wu, Yunpu Ma, Ruotong Liao, Bo Xiong, and Volker Tresp. zrllm: Zero-shot relational learning on temporal knowledge graphs with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1877–1895, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

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

Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. Automatically labeling clinical trial outcomes: A large-scale benchmark for drug development. *arXiv preprint arXiv:2406.10292*, 2024.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event prediction. *ACM Transactions on Information Systems*, 2024.

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

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. LightRAG: Simple and Fast Retrieval-Augmented Generation. In *EMNLP (Findings)*, pp. 10746–10761, 2025.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chinmay Karkar and Paras Chopra. Future is unevenly distributed: Forecasting ability of llms depends on what we're asking. *arXiv preprint arXiv:2511.18394*, 2025.

Koray Kavukcuoglu. Gemini 2.0 is now available to everyone. Google DeepMind, February 2025. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-updates-february-2025/.

Tom A Lamb, Desi R Ivanova, Philip Torr, and Tim GJ Rudner. Semantic-level confidence calibration of language models via temperature scaling. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 544–557, 2023.

Sang-Woo Lee, Sohee Yang, Donghyun Kwak, and Noah Y Siegel. Advancing event forecasting through massive training of large language models: Challenges, solutions, and broader impacts. *arXiv preprint arXiv:2507.19477*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Ningyuan Li, E Haihong, Tianyu Yao, Tianyi Hu, Yuhan Li, Haoran Luo, Meina Song, and Yifan Zhu. Infer: a neural-symbolic model for extrapolation reasoning on temporal knowledge graph. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 4303–4317, 2024.

LightningRodLabs. LightningRodLabs/ai-futures-2025. `https://huggingface.co/datasets/LightningRodLabs/ai-futures-2025`, 2025a. Hugging Face dataset.

LightningRodLabs. LightningRodLabs/GolfForecasting. `https://huggingface.co/datasets/LightningRodLabs/GolfForecasting`, 2025b. Hugging Face dataset.

Janna Lu. Evaluating llms on real-world forecasting against expert forecasters. *arXiv preprint arXiv:2507.04562*, 2025.

Yuyin Lu, Ziran Liang, Yanghui Rao, Wenqi Fan, Fu Lee Wang, and Qing Li. Double-calibration: Towards trustworthy llms via calibrating knowledge and reasoning confidence. *arXiv preprint arXiv:2601.11956*, 2026.

Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, et al. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *arXiv preprint arXiv:2503.21322*, 2025.

Kun Luo, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. Open event causality extraction by the assistance of llm in task annotation, dataset, and method. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge)@ LREC-COLING-2024*, pp. 33–44, 2024.

Putra Manggala, Atalanti Mastakouri, Elke Kirschbaum, Shiva Prasad Kasiviswanathan, and Aaditya Ramdas. Qa-calibration of language model confidence scores. *arXiv preprint arXiv:2410.06615*, 2024.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in llms. *arXiv preprint arXiv:2511.04869*, 2025.

Lukas Nel. Do large language models know what they don't know? kalshibench: A new benchmark for evaluating epistemic calibration via prediction markets, 2025. URL `https://arxiv.org/abs/2512.16030`.

Luyao Niu, Zepu Wang, Shuyi Guan, Yang Liu, and Peng Sun. Event-causnet: Unlocking causal knowledge from text with large language models for reliable spatio-temporal forecasting. *arXiv preprint arXiv:2511.12769*, 2025.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. *arXiv preprint arXiv:2412.18544*, 2024.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723*, 2025.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52, 2025.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Polymarket. Polymarket: Information markets. `https://polymarket.com`, 2024. Dataset snapshot available at `https://huggingface.co/datasets/CK0607/polymarket_10000`.

Jing Ren, Bowen Li, Ziqi Xu, Xikun Zhang, Haytham Fayek, and Xiaodong Li. When to trust: A causality-aware calibration framework for accurate knowledge graph retrieval-augmented generation. *arXiv preprint arXiv:2601.09241*, 2026.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Zhengwei Tao, Pu Wu, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Xiancai Chen, Jia Li, Linyu Li, Chongyang Tao, et al. Prophet: An inferable future forecasting benchmark with causal intervened likelihood estimation. *arXiv preprint arXiv:2504.01509*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Benjamin Turtel, Paul Wilczewski, Danny Franklin, and Kris Skothiem. Future-as-label: Scalable supervision from real-world outcomes. *arXiv preprint arXiv:2601.06336*, 2026.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024a.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Beyond inherent cognition biases in llm-based event forecasting: A multi-cognition agentic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4799–4818, 2025.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jack Wildman, Nikos I Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, Lawrence Phillips, et al. Bench to the future: A pastcasting benchmark for forecasting agents. *arXiv preprint arXiv:2506.21558*, 2025.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. *arXiv preprint arXiv:2505.01997*, 2025a.

Yilin Xiao, Junnan Dong, Chuang Zhou, Su Dong, Qian-wen Zhang, Di Yin, Xing Sun, and Xiao Huang. Graphrag-bench: Challenging domain-specific reasoning for evaluating graph retrieval-augmented generation. *arXiv preprint arXiv:2506.02404*, 2025b.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, 2024a.

Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. Calibrating reasoning in language models with internal consistency. *Advances in Neural Information Processing Systems*, 37:114872–114901, 2024b.

Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. *arXiv preprint arXiv:2310.01880*, 2023.

Hanchen Yang, Jiaqi Wang, Jiannong Cao, Wengen Li, Jialun Zheng, Yangning Li, Chunyu Miao, Jihong Guan, Shuigeng Zhou, and Philip S Yu. Okg-llm: Aligning ocean knowledge graph with observation data via llms for global sea surface temperature prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Forecast: The future outcome reasoning and confidence assessment benchmark. *arXiv preprint arXiv:2502.19676*, 2025.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pp. 609–616, 2001.

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*, pp. 694–699, 2002.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, et al. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*, 2025.

Meiru Zhang, Auss Abbood, Zaiqiao Meng, and Nigel Collier. The power of simplicity in llm-based event forecasting. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 454–470, 2025.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, 2024.

Yinan Zhang, Zhixi Chen, Jiazheng Jing, and Zhiqi Shen. Following the trail: Predicting and explaining tomorrow's hits with a fine-tuned llm. *arXiv preprint arXiv:2602.04225*, 2026.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35:27293–27305, 2022.

APPENDIX

A THEORETICAL PROOF

A.1 PROOF OF PROPOSITION 1

Proposition 1. *Under normalized weighting, CTVF lowers weighting error versus uniform weighting by favoring lower-error proximal evidence without increasing the weighted error within either subset.*

Proof. To isolate the weighting effect of CTVF from subsequent chain aggregation, let $\mathcal{E} = \{e_i\}_{i=1}^{M}$ denote the candidate evidence set, where each e_i carries an effective causal topological distance $d_i \in \{0, 1, \ldots, d_{\max}\}$, obtained by capping its raw distance at $d_{\max}$ and assigning unreachable evidence $d_i = d_{\max}$. Let $V_i > 0$ be the validity assigned by CTVF, $p_i \in [0, 1]$ the per-evidence event probability, $y \in \{0, 1\}$ the binary ground truth, and $\xi_i = |p_i - y|$ the corresponding error term.

Fix a threshold $d^* \in \{0, 1, \ldots, d_{\max} - 1\}$ for which the causally proximal set $\mathcal{P} = \{i : d_i \le d^*\}$ and the causally distal set $\mathcal{D} = \{i : d_i > d^*\}$ are both nonempty. Let $M_{\mathcal{P}} = |\mathcal{P}| > 0$ and $M_{\mathcal{D}} = |\mathcal{D}| > 0$, so that

$$M = M_{\mathcal{P}} + M_{\mathcal{D}}. \tag{23}$$

For any weighting scheme $\Lambda = \{\lambda_i\}_{i=1}^{M}$ satisfying $\sum_{i=1}^{M} \lambda_i = 1$ and $\lambda_i \ge 0$, define the weighting error as

$$\mathcal{E}(\Lambda) = \sum_{i=1}^{M} \lambda_i \xi_i. \tag{24}$$

Under uniform weighting,

$$\lambda_i^{\text{uni}} = \frac{1}{M}, \tag{25}$$

and hence

$$\mathcal{E}(\Lambda^{\text{uni}}) = \frac{1}{M} \sum_{i=1}^{M} \xi_i = \frac{M_{\mathcal{P}} \bar{\xi}_{\mathcal{P}} + M_{\mathcal{D}} \bar{\xi}_{\mathcal{D}}}{M}, \tag{26}$$

where

$$\bar{\xi}_{\mathcal{P}} = \frac{1}{M_{\mathcal{P}}} \sum_{i \in \mathcal{P}} \xi_i, \qquad \bar{\xi}_{\mathcal{D}} = \frac{1}{M_{\mathcal{D}}} \sum_{i \in \mathcal{D}} \xi_i \tag{27}$$

denote the mean errors of the proximal and distal subsets, respectively.

Under CTVF weighting,

$$\lambda_i^{\text{CTVF}} = \frac{V_i}{C_V}, \qquad C_V = \sum_{j=1}^{M} V_j. \tag{28}$$

Define the mean validities of the two subsets as

$$\bar{V}_{\mathcal{P}} = \frac{1}{M_{\mathcal{P}}} \sum_{i \in \mathcal{P}} V_i, \qquad \bar{V}_{\mathcal{D}} = \frac{1}{M_{\mathcal{D}}} \sum_{i \in \mathcal{D}} V_i, \tag{29}$$

and their total validities as

$$S_{\mathcal{P}} = M_{\mathcal{P}} \bar{V}_{\mathcal{P}}, \qquad S_{\mathcal{D}} = M_{\mathcal{D}} \bar{V}_{\mathcal{D}}. \tag{30}$$

Further define the within-subset CTVF-weighted mean errors by

$$\tilde{\xi}_{\mathcal{P}} = \frac{\sum_{i \in \mathcal{P}} V_i \xi_i}{\sum_{i \in \mathcal{P}} V_i}, \qquad \tilde{\xi}_{\mathcal{D}} = \frac{\sum_{i \in \mathcal{D}} V_i \xi_i}{\sum_{i \in \mathcal{D}} V_i}. \tag{31}$$

Then the CTVF weighting error can be written exactly as

$$\mathcal{E}(\Lambda^{\text{CTVF}}) = \frac{S_{\mathcal{P}} \tilde{\xi}_{\mathcal{P}} + S_{\mathcal{D}} \tilde{\xi}_{\mathcal{D}}}{S_{\mathcal{P}} + S_{\mathcal{D}}}. \tag{32}$$

Define the total weight mass assigned to the proximal subset under the two schemes as

$$\rho^* = \frac{S_{\mathcal{P}}}{S_{\mathcal{P}} + S_{\mathcal{D}}}, \qquad \rho_0 = \frac{M_{\mathcal{P}}}{M}. \tag{33}$$

Accordingly,

$$\mathcal{E}(\Lambda^{\mathrm{uni}}) = \rho_0 \bar{\xi}_{\mathcal{P}} + (1-\rho_0)\bar{\xi}_{\mathcal{D}}, \tag{34}$$

and

$$\mathcal{E}(\Lambda^{\mathrm{CTVF}}) = \rho^* \tilde{\xi}_{\mathcal{P}} + (1-\rho^*)\tilde{\xi}_{\mathcal{D}}. \tag{35}$$

Their difference admits the exact decomposition

$$\mathcal{E}(\Lambda^{\mathrm{uni}}) - \mathcal{E}(\Lambda^{\mathrm{CTVF}}) = \big[\rho_0(\bar{\xi}_{\mathcal{P}} - \tilde{\xi}_{\mathcal{P}}) + (1-\rho_0)(\bar{\xi}_{\mathcal{D}} - \tilde{\xi}_{\mathcal{D}})\big] + (\rho^* - \rho_0)(\tilde{\xi}_{\mathcal{D}} - \tilde{\xi}_{\mathcal{P}}). \tag{36}$$

This identity separates the error gap into two components. The first term measures the within-subset reweighting effect, namely whether CTVF concentrates more mass on lower-error evidence inside each subset. The second term measures the between-subset reallocation effect, namely whether CTVF assigns more total mass to the proximal subset while that subset remains the lower-error one after within-subset weighting.

Therefore, whenever the CTVF-induced reweighting relations satisfy

$$\tilde{\xi}_{\mathcal{P}} \le \bar{\xi}_{\mathcal{P}}, \qquad \tilde{\xi}_{\mathcal{D}} \le \bar{\xi}_{\mathcal{D}}, \tag{37}$$

and

$$\rho^* > \rho_0, \qquad \tilde{\xi}_{\mathcal{P}} < \tilde{\xi}_{\mathcal{D}}, \tag{38}$$

the first bracket is non-negative and the second term is strictly positive. Consequently,

$$\mathcal{E}(\Lambda^{\mathrm{CTVF}}) < \mathcal{E}(\Lambda^{\mathrm{uni}}). \tag{39}$$

Therefore, whenever the above CTVF-induced reweighting relations hold, CTVF reduces weighting error along two complementary channels: it concentrates weight on lower-error evidence inside each causal subset through validity differences, and reallocates higher total mass to the causally proximal subset across subsets. Under these causal-proximity and error-ordering conditions, the normalized CTVF weights consequently achieve a strictly lower weighting error than uniform weights. □

A.2 Proof of Proposition 2

Proposition 2. *Direction-aware deduplication can eliminate the same-polarity Noisy-OR probability inflation induced by overlap-identified redundant causal chains.*

Proof. Fix a question q, let Π denote either same-polarity set Π_q^+ or Π_q^-, and let $\widehat{\Pi} \subseteq \Pi$ be its deduplicated counterpart, for which any two retained chains have Jaccard overlap at most θ and every discarded chain $\pi \in \Pi \setminus \widehat{\Pi}$ admits some $\pi' \in \widehat{\Pi}$ with $J(\pi, \pi') > \theta$. The same-polarity Noisy-OR aggregation $p_{\mathrm{NoisyOR}}(\Pi) = 1 - \prod_{\pi\in\Pi}(1 - \Phi(\pi, q))$ is the closed-form union probability under the Noisy-OR independence assumption. Treating an overlap-identified redundant chain as an additional independent factor therefore weakly inflates the computed same-polarity aggregate.

Splitting the Noisy-OR product over Π between $\widehat{\Pi}$ and $\Pi \setminus \widehat{\Pi}$:

$$\prod_{\pi\in\Pi}\big(1 - \Phi(\pi, q)\big) = \prod_{\pi\in\widehat{\Pi}}\big(1 - \Phi(\pi, q)\big) \cdot \prod_{\pi\in\Pi\setminus\widehat{\Pi}}\big(1 - \Phi(\pi, q)\big) \le \prod_{\pi\in\widehat{\Pi}}\big(1 - \Phi(\pi, q)\big), \tag{40}$$

since $\Phi(\pi, q) \in [0, 1]$ keeps the second factor in $[0, 1]$. Hence

$$p_{\mathrm{NoisyOR}}(\Pi) \ge p_{\mathrm{NoisyOR}}(\widehat{\Pi}), \tag{41}$$

and the inequality is strict whenever $p_{\mathrm{NoisyOR}}(\widehat{\Pi}) < 1$ and a removed chain has $\Phi(\pi, q) > 0$.

Define the polarity-wise redundancy factors

$$\eta^+ = \prod_{\pi\in\Pi_q^+\setminus\widehat{\Pi}_q^+}\big(1 - \Phi(\pi, q)\big) \in [0, 1], \qquad \eta^- = \prod_{\pi\in\Pi_q^-\setminus\widehat{\Pi}_q^-}\big(1 - \Phi(\pi, q)\big) \in [0, 1], \tag{42}$$

and the naive pre-deduplication probabilities $p^{\pm}_{\text{naive}} = p_{\text{NoisyOR}}(\Pi^{\pm}_q)$. The splitting identity directly gives

$$1 - p^{+}_{\text{naive}} = \eta^{+}\big(1 - p^{+}(q)\big), \qquad 1 - p^{-}_{\text{naive}} = \eta^{-}\big(1 - p^{-}(q)\big). \tag{43}$$

Applying Boole's inequality $\mathbb{P}(\bigcup A_\pi) \le \sum_\pi \mathbb{P}(A_\pi)$ to the duplicated chains, equivalently the Weierstrass product inequality $\prod(1-\phi_i) \ge 1 - \sum \phi_i$ for $\phi_i \in [0,1]$, yields

$$1 - \eta^{\pm} \;\le\; B^{\pm}, \qquad B^{\pm} \;:=\; \sum_{\pi \in \Pi^{\pm}_q \setminus \widehat{\Pi}^{\pm}_q} \Phi(\pi, q), \tag{44}$$

and combining with $1 - p^{\pm}_{\text{naive}} = \eta^{\pm}(1 - p^{\pm})$ delivers the polarity-wise inflation upper bound

$$p^{\pm}_{\text{naive}} - p^{\pm}(q) = (1 - p^{\pm}(q)) \cdot (1 - \eta^{\pm}) \;\le\; (1 - p^{\pm}(q)) \cdot B^{\pm} \;\le\; B^{\pm}, \tag{45}$$

which characterizes the maximum probability-level deviation that omitting deduplication can incur and depends only on the confidences of the removed redundant chains. Whenever $\eta^{\pm} < 1$, $p^{\pm}_{\text{naive}} \ge p^{\pm}(q)$; when $p^{\pm}(q) = 1$, $p^{\pm}_{\text{naive}} = 1$ and the inflation vanishes.

Greedy deduplication retains at least one chain from every non-empty polarity. Define the neutral-anchored transformation

$$g_\varepsilon(p) = \operatorname{logit}\big(\min(1-\varepsilon, \max(\varepsilon, (1+p)/2))\big). \tag{46}$$

Applying the same unified combination before deduplication gives

$$P_{\text{naive}}(q) = \sigma\big(g_\varepsilon(p^{+}_{\text{naive}}) - g_\varepsilon(p^{-}_{\text{naive}})\big). \tag{47}$$

Because g_ε and σ are non-decreasing, redundancy confined to the positive polarity gives $P_{\text{naive}}(q) \ge P_{\text{causal}}(q)$, whereas redundancy confined to the negative polarity gives $P_{\text{naive}}(q) \le P_{\text{causal}}(q)$; each inequality is strict whenever a removed positive-score chain changes the affected clipped-logit value. When both polarities contain redundancy, both polarity-wise Noisy-OR scores are inflated, while the direction of the final change depends on their relative increments. The same monotonicity argument applies without separate single-channel branches, and two empty or equally strong channels return $\frac{1}{2}$.

Therefore, direction-aware deduplication removes the overlap-identified factors responsible for the bounded same-polarity Noisy-OR inflation, while the neutral-anchored mapping preserves the direction of each polarity-specific change in the final causal probability. □

A.3 Proof of Proposition 3

Proposition 3. *Adaptive fusion using causal coverage and directional balance down-weights the causal estimate when evidence is unreliable, and can yield lower squared error than fixed weighting.*

Proof. Let q be drawn from a distribution on $\mathcal{Q}$, and let $P_{\text{causal}}(q), P_{\text{llm}}(q) \in [0,1]$ denote two probability estimates of the true probability $p^\star(q) \in [0,1]$. Define the corresponding estimation errors by

$$\epsilon_C(q) = P_{\text{causal}}(q) - p^\star(q), \qquad \epsilon_L(q) = P_{\text{llm}}(q) - p^\star(q). \tag{48}$$

For any weighting function $\alpha : \mathcal{Q} \to [0, \alpha_0]$, the fused forecast is

$$\hat{p}_\alpha(q) = \alpha(q) P_{\text{causal}}(q) + \big(1 - \alpha(q)\big) P_{\text{llm}}(q), \tag{49}$$

with fused error

$$\hat{p}_\alpha(q) - p^\star(q) = \alpha(q)\epsilon_C(q) + (1 - \alpha(q))\epsilon_L(q). \tag{50}$$

Define the corresponding squared-error risk by

$$R(\alpha) = \mathbb{E}_q\Big[\big(\alpha(q)\epsilon_C(q) + (1 - \alpha(q))\epsilon_L(q)\big)^2\Big]. \tag{51}$$

For $\alpha_b > 0$, $\zeta > 0$, $0 < \alpha_0 \le 1$, and $\alpha_0 < 2\alpha_b$, define the saturation threshold $\Omega_{\text{sat}} := \Omega_0 + \zeta^{-1} \operatorname{logit}(\alpha_0/(2\alpha_b))$ and the partition

$$\mathcal{G}_{\text{hi}} = \{q : k(q) > 0,\ \Omega(q) \ge \Omega_{\text{sat}}\}, \qquad \mathcal{G}_{\text{lo}} = \mathcal{Q} \setminus \mathcal{G}_{\text{hi}}, \tag{52}$$

and denote

$$\pi_g := \mathbb{P}(\mathcal{G}_g), \qquad g \in \{\text{hi}, \text{lo}\}, \qquad \pi_{\text{hi}} + \pi_{\text{lo}} = 1. \tag{53}$$

Assume $0 < \pi_{\mathrm{hi}}, \pi_{\mathrm{lo}} < 1$, so that both conditional distributions are well defined. Conditional expectations below are taken with respect to this partition.

Suppose the causal estimate is less reliable on the low-reliability stratum: there exists $\Delta > 0$ such that

$$\mathbb{E}\big[\epsilon_C^2(q) \mid \mathcal{G}_{\mathrm{lo}}\big] - \mathbb{E}\big[\epsilon_C^2(q) \mid \mathcal{G}_{\mathrm{hi}}\big] \geq \Delta. \tag{54}$$

Suppose further that the LLM-side error structure is stratum-invariant: there exist constants B and C with

$$\mathbb{E}\big[\epsilon_L^2(q) \mid \mathcal{G}_g\big] = B, \qquad \mathbb{E}[\epsilon_C(q)\epsilon_L(q) \mid \mathcal{G}_g] = C, \qquad g \in \{\mathrm{hi}, \mathrm{lo}\}. \tag{55}$$

Suppose finally that $B > C$, expressing non-degeneracy of the LLM contribution.

By the tower property of conditional expectation,

$$R(\alpha) = \sum_{g \in \{\mathrm{hi},\mathrm{lo}\}} \pi_g \, \mathbb{E}\Big[\big(\alpha(q)\epsilon_C + (1-\alpha(q))\epsilon_L\big)^2 \,\Big|\, \mathcal{G}_g\Big]. \tag{56}$$

For any weighting that takes a constant value α on $\mathcal{G}_g$, the conditional risk is

$$R_g^{\mathrm{cst}}(\alpha) = \alpha^2 A_g + (1-\alpha)^2 B + 2\alpha(1-\alpha)C, \qquad A_g := \mathbb{E}[\epsilon_C^2 \mid \mathcal{G}_g]. \tag{57}$$

The quadratic coefficient admits the identity

$$A_g + B - 2C = \mathbb{E}\big[(\epsilon_C - \epsilon_L)^2 \mid \mathcal{G}_g\big] \geq 0. \tag{58}$$

Strict positivity follows from the Cauchy-Schwarz bound $C^2 \leq A_g B$ together with $B > C$: when $C \geq 0$, $A_g + B - 2C \geq (B-C)^2/B > 0$; when $C < 0$, $A_g + B - 2C > A_g + B > 0$. The coefficient is therefore strictly positive in both regimes, so $R_g^{\mathrm{cst}}(\alpha)$ is a strictly convex quadratic in α with unique unconstrained minimizer

$$\alpha_g^* = \frac{B - C}{A_g + B - 2C}, \tag{59}$$

and feasible minimizer

$$\bar{\alpha}_g^* = \mathrm{clip}_{[0,\alpha_0]}(\alpha_g^*). \tag{60}$$

The two stratum-wise optima satisfy

$$\alpha_{\mathrm{hi}}^* - \alpha_{\mathrm{lo}}^* = \frac{(B-C)(A_{\mathrm{lo}} - A_{\mathrm{hi}})}{(A_{\mathrm{hi}} + B - 2C)(A_{\mathrm{lo}} + B - 2C)}. \tag{61}$$

Since $B - C > 0$ and

$$A_{\mathrm{lo}} - A_{\mathrm{hi}} \geq \Delta > 0, \tag{62}$$

it follows that

$$\alpha_{\mathrm{hi}}^* > \alpha_{\mathrm{lo}}^*, \qquad \bar{\alpha}_{\mathrm{hi}}^* \geq \bar{\alpha}_{\mathrm{lo}}^*. \tag{63}$$

Hence the constrained optimal fusion rule assigns no less weight to the causal estimate in the high-reliability stratum than in the low-reliability stratum.

For any fixed-weight scheme $\alpha_{\mathrm{fix}} \in [0, \alpha_0]$, the same constant is assigned on both strata, so

$$R(\alpha_{\mathrm{fix}}) = \pi_{\mathrm{hi}} R_{\mathrm{hi}}^{\mathrm{cst}}(\alpha_{\mathrm{fix}}) + \pi_{\mathrm{lo}} R_{\mathrm{lo}}^{\mathrm{cst}}(\alpha_{\mathrm{fix}}). \tag{64}$$

Strict convexity gives

$$R_g^{\mathrm{cst}}(\alpha_{\mathrm{fix}}) \geq R_g^{\mathrm{cst}}(\bar{\alpha}_g^*), \tag{65}$$

hence

$$R(\alpha_{\mathrm{fix}}) - \Big[\pi_{\mathrm{hi}} R_{\mathrm{hi}}^{\mathrm{cst}}(\bar{\alpha}_{\mathrm{hi}}^*) + \pi_{\mathrm{lo}} R_{\mathrm{lo}}^{\mathrm{cst}}(\bar{\alpha}_{\mathrm{lo}}^*)\Big] = \sum_g \pi_g \Big[R_g^{\mathrm{cst}}(\alpha_{\mathrm{fix}}) - R_g^{\mathrm{cst}}(\bar{\alpha}_g^*)\Big] \geq 0. \tag{66}$$

If $\bar{\alpha}_{\mathrm{hi}}^* \neq \bar{\alpha}_{\mathrm{lo}}^*$, then α_{fix} cannot simultaneously attain both stratum-wise optima, so at least one stratum is strictly suboptimal, and

$$R(\alpha_{\mathrm{fix}}) > \pi_{\mathrm{hi}} R_{\mathrm{hi}}^{\mathrm{cst}}(\bar{\alpha}_{\mathrm{hi}}^*) + \pi_{\mathrm{lo}} R_{\mathrm{lo}}^{\mathrm{cst}}(\bar{\alpha}_{\mathrm{lo}}^*). \tag{67}$$

Define the adaptive weighting rule by

$$\alpha(q) := \mathbb{1}[k(q) > 0]\ \min\big(2\alpha_b\sigma(\zeta(\Omega(q) - \Omega_0)), \alpha_0\big). \tag{68}$$

Since $\zeta > 0$ and σ is strictly increasing, $q \in \mathcal{G}_{\mathrm{hi}}$ gives $k(q) > 0$ and $\Omega(q) \geq \Omega_{\mathrm{sat}}$, hence

$$2\alpha_b\sigma(\zeta(\Omega(q) - \Omega_0)) \geq \alpha_0, \tag{69}$$

and the truncation gives

$$\alpha(q) = \alpha_0. \tag{70}$$

For $q \in \mathcal{G}_{\mathrm{lo}}$, either $k(q) = 0$, which gives $\alpha(q) = 0$, or $\Omega(q) < \Omega_{\mathrm{sat}}$, which gives

$$\alpha(q) < \alpha_0. \tag{71}$$

Denote $\bar{\alpha}_g := \mathbb{E}[\alpha(q) \mid \mathcal{G}_g]$. Then

$$\bar{\alpha}_{\mathrm{hi}} = \alpha_0 > \bar{\alpha}_{\mathrm{lo}}, \tag{72}$$

ordered in the same direction as $\bar{\alpha}^*_{\mathrm{hi}} \geq \bar{\alpha}^*_{\mathrm{lo}}$.

Assume additionally that, within each stratum, conditioning on $\alpha(q)$ leaves the three conditional moments A_g, B, and C defined above unchanged. Since $\alpha(q)$ is constant on $\mathcal{G}_{\mathrm{hi}}$ but may vary on $\mathcal{G}_{\mathrm{lo}}$, the adaptive risk admits a within-stratum variance term. As R^{cst}_g is quadratic, its exact second-order Taylor expansion at $\bar{\alpha}_g$ gives

$$\mathbb{E}[R^{\mathrm{cst}}_g(\alpha(q)) \mid \mathcal{G}_g] = R^{\mathrm{cst}}_g(\bar{\alpha}_g) + (A_g + B - 2C)\,\mathrm{Var}(\alpha(q) \mid \mathcal{G}_g), \tag{73}$$

hence

$$R(\alpha) = \sum_g \pi_g \Big[R^{\mathrm{cst}}_g(\bar{\alpha}_g) + (A_g + B - 2C)\,\mathrm{Var}(\alpha(q) \mid \mathcal{G}_g)\Big], \tag{74}$$

with

$$\mathrm{Var}(\alpha(q) \mid \mathcal{G}_{\mathrm{hi}}) = 0, \qquad \mathrm{Var}(\alpha(q) \mid \mathcal{G}_{\mathrm{lo}}) \geq 0. \tag{75}$$

Therefore, for any fixed-weight scheme $\alpha_{\mathrm{fix}} \in [0, \alpha_0]$, if

$$\sum_g \pi_g \big[R^{\mathrm{cst}}_g(\alpha_{\mathrm{fix}}) - R^{\mathrm{cst}}_g(\bar{\alpha}_g)\big] > \sum_g \pi_g (A_g + B - 2C)\,\mathrm{Var}(\alpha(q) \mid \mathcal{G}_g), \tag{76}$$

it follows that

$$R(\alpha) < R(\alpha_{\mathrm{fix}}) \tag{77}$$

The directional-balance factor $\widetilde{\beta}(q)$ enters $\alpha(q)$ through the coverage score $\Omega(q)$: a smaller $\widetilde{\beta}(q)$ cannot increase $\Omega(q)$ and therefore cannot increase $\alpha(q)$.

Hence the adaptive fusion weight $\alpha(q)$ equals α_0 on $\mathcal{G}_{\mathrm{hi}}$ and is strictly smaller than α_0 on $\mathcal{G}_{\mathrm{lo}}$. Whenever the weighted reduction in stratum-wise mean risk exceeds the within-stratum variance penalty, the strict inequality $R(\alpha) < R(\alpha_{\mathrm{fix}})$ shows that adaptive fusion can attain lower squared-error risk than fixed weighting. □

B Prompts Used in CHAIN

B.1 Hypergraph and Causal Extraction Prompt

This prompt drives the construction of the causal-temporal hypergraph $\mathcal{G} = (V, E_H, \mathcal{C})$ that underlies **CHAIN**. Invoked once per factual chunk, it implements the LLM extractor ϕ_{LLM} that jointly produces hyperedge-level factual propositions (forming E_H) and typed directed causal edges with strength scores (forming $\mathcal{C}$). The prompt is shown in Figure 7.

[SYSTEM]

You extract one strict causal-temporal JSON object and must perform the final RFC 8259 escaping check before returning it. Every literal backslash in a JSON string must be encoded as two consecutive backslash characters; never emit an unsupported escape or raw LaTeX command. Also perform the prompt's final entity-count check immediately before returning it. Never return a proposition whose entities array is empty; if any entity in a proposition is not an exact substring of the supplied text, omit the entire proposition rather than deleting or repairing only that entity. Every admitted causal relation must have a strictly positive textual-support strength; never emit zero. The only permitted causal relation type values are causes, enables, and prevents. Omit any relation that cannot use one of those exact values; never invent another type string.

[USER]

Extract factual propositions and causal relationships from the supplied text.

Return exactly one JSON object with exactly "propositions" and "causal_relations" arrays. Do not output any other top-level field.

The response MUST be one RFC 8259 JSON value. Every string value MUST use valid JSON escaping. A literal backslash from the supplied Text MUST be encoded as two consecutive backslash characters in the JSON encoding; never emit an unsupported escape such as a single backslash followed by a letter, underscore, or dollar sign. Do not emit raw LaTeX commands inside JSON strings. Use plain Unicode/text when possible; when a verbatim source phrase contains a backslash, preserve it only with valid JSON escaping. After JSON decoding, the resulting string value must still satisfy every verbatim grounding rule below. Never use a single JSON escape such as \b, \f, \n, \r, or \t to represent the beginning of a source LaTeX command: those escapes decode to control characters rather than a literal backslash. Every decoded string must contain no XML 1.0-forbidden control character. Check the complete response with a strict JSON parser before returning it.

Every proposition object MUST contain exactly:
{"sentence":"non-empty factual sentence","entities":[...]}

Do not output any other proposition field. Every entity object MUST contain either exactly {"name":"non-empty entity name"} or exactly {"name":"non-empty entity name","name_zh":"optional Chinese name"}. Do not output entity types, IDs, descriptions, aliases, confidence, or any other entity field. Every emitted entity "name" MUST copy, byte-for-byte, one complete contiguous substring from the supplied Text, including its exact capitalization, punctuation, Unicode form, and internal whitespace. The "entities" array MUST be non-empty. If ANY emitted entity in a proposition cannot be grounded this way, OMIT THAT ENTIRE PROPOSITION; never return an empty "entities" array, delete only the bad entity, or invent, repair, normalize, paraphrase, translate, or complete an entity from a truncated or incomplete sentence.

ENTITY SURFACE-FORM RULE: when the same entity appears in multiple propositions, reuse one consistent spelling, capitalization, Unicode form, and whitespace pattern for its name. Do not emit separate entity names that differ only under case-folding, Unicode compatibility normalization, or whitespace normalization. The builder retains any such surface forms as aliases under one canonical entity identity; this does not merge names with different canonical keys. Within one proposition, list each canonical entity only once. If two distinct referents could otherwise collapse under this normalization, use explicit disambiguating names rather than case alone.

Every causal relation object MUST contain exactly the required fields "cause", "effect", "type", and "strength", and MAY also contain the single optional string field "description". Do not output any other relation field. The "strength" value is a finite JSON number strictly in (0,1]. Estimate it independently for each relation from the supplied text; do not copy a fixed number from these instructions or reuse one default value across relations.

CRITICAL ENDPOINT IDENTITY RULE: treat the emitted propositions[*].entities[*].name strings as a closed vocabulary. Every cause/effect MUST copy, byte-for-byte, one complete name string from that vocabulary in this same response. Never use a complete sentence, pronoun, summary, description, alias, normalized spelling, or paraphrase as an endpoint. If you cannot confirm an exact byte-for-byte match, OMIT the causal relation; do not invent, repair, shorten, or rewrite an endpoint. If a compound event is needed, first emit it as an entity name and then copy that exact name.

The supplied Text may end in the middle of a word or sentence. Never complete, repair, or infer missing text after a truncation; omit any fact or causal relation that cannot be supported without the missing continuation. Before returning JSON, self-check every causal relation by locating both endpoint strings in the emitted entity names and confirming exact equality, including capitalization, punctuation, and whitespace. The canonicalized cause and effect MUST be different; otherwise OMIT the self-loop relation. Do not repeat one canonical (cause,effect,type) relation through alternate entity surface forms.

<...example...>

Relation type is exactly one of "causes", "enables", and "prevents". The literal JSON string after every "type" key MUST match exactly one member of that closed vocabulary. If a relationship cannot use one of those exact values, OMIT the causal relation entirely; do not invent, map, normalize, or coerce a type. Before returning, scan every relation and confirm no other type string occurs.

Strength is a finite JSON number strictly in (0,1] and must not be rounded. It is the positive textual causal-support score for this edge, not the forecast probability and not a default placeholder. Choose a value reflecting how directly the supplied text supports the typed relation. If the text does not support a positive causal relation, omit that relation; never emit strength 0 or 0.0. Do not repeat a proposition or a (cause,effect,type) relation. If no facts exist return:
{"propositions":[],"causal_relations":[]}

FINAL ENTITY-COUNT CHECK: For every proposition, count its entities. If the count would be zero, delete the entire proposition. Never output an empty entities array. A generic event or date sentence is admissible only if you emit at least one exact verbatim noun/event/date phrase from that sentence as an entity; otherwise omit the proposition. Do not retain any proposition with "entities": [].

Text:
{text}

Figure 7: Hypergraph and Causal Extraction Prompt.

[SYSTEM]

You are an expert forecasting analyst. Predict whether the specified future event will occur using only the supplied evidence.

[USER]

You are resolving causal targets for a probabilistic forecast.

Question: {question}
Prediction cutoff: {cutoff_date}
Only evidence strictly admissible before this cutoff may be used.

Canonical binary axis:
- positive/event: id={positive_id}; label={positive_label}; semantics={positive_semantics}
- negative/complement: id={negative_id}; label={negative_label}; semantics={negative_semantics}

Admitted graph entity labels:
{graph_summary}

Choose only unambiguous question-resolving target anchors. Copy each entity label exactly from the admitted list. Mark axis_side as positive when an increase or occurrence supports the event side, and negative when it supports the complement side. If no anchor is unambiguous, return an empty list.

Return exactly this JSON schema:
```
{
 "entities": ["relevant admitted entity"],
 "target_anchors": [
   {
    "entity": "exact admitted label",
    "axis_side": "positive"
   }
 ],
 "hypotheses": ["causal hypothesis"],
 "dimensions": ["forecast dimension"],
 "search_queries": ["query"],
 "confidence_prior": 0.5
}
```

Return exactly one valid JSON object. Do not use markdown fences, comments, trailing text, NaN, or Infinity.

Figure 8: Reasoning Direction Generation Prompt.

```
[SYSTEM]

You are an expert forecasting analyst. Predict whether the specified future event will occur using only the supplied evidence.

[USER]

You are a probabilistic future-event prediction agent.

Round: {round_num}
Question: {question}
Prediction cutoff: {cutoff_date}
Canonical positive/event side: {positive_label} — {positive_semantics}
Canonical negative/complement side: {negative_label} — {negative_semantics}

All evidence below comes from one cutoff-safe active graph view. Known evidence at or after the prediction cutoff is excluded. Missing event time is included only with a verified availability upper bound strictly before the cutoff. Scientific evidence weights use the Causal-Temporal Validity Function (CTVF).

CONTEXT 1 — ENTITY NARRATIVES
{narratives}

CONTEXT 2 — OVERALL TRAJECTORY
{trajectory}

CONTEXT 3 — CONTRADICTIONS
{contradictions}

CONTEXT 4 — LIGHTWEIGHT CAUSAL-CONTEXT CHAINS
{causal_context}

CONTEXT 5 — SOURCE EVENTS
{source_events}

PRIOR ROUND OUTCOMES
{prior_outcomes}

Define mutually exclusive, collectively exhaustive outcomes that directly answer the question at its resolution horizon. Use causal mechanisms, CTVF-weighted evidence, trajectory, contradictions, and structural constraints. Do not describe merely the current trend. Probabilities must be finite, non-negative, and sum to exactly 1 within the requested numerical tolerance.

Return exactly this JSON schema:
{
 "outcomes": [
   {
    "id": "provider-local optional id",
    "name": "concise resolved future state",
    "description": "one sentence world state at resolution",
    "indicators": ["evidence type"],
    "probability": 0.25
   }
 ],
 "pivot_factor": "single decisive variable",
 "need_more_info": false,
 "info_queries": [],
 "reasoning": "concise causal explanation"
}

Return exactly one valid JSON object. Do not use markdown fences, comments, trailing text, NaN, or Infinity.
```

Figure 9: Multi-Outcome Probabilistic Reasoning Prompt.

```
[SYSTEM]

You are an expert forecasting analyst. Predict whether the specified future event will occur using only the supplied evidence.

[USER]

Map every forecast outcome to exactly one side of the canonical binary axis.

Question: {question}
Prediction cutoff: {cutoff_date}
Positive/event side: id={positive_id}; label={positive_label}; semantics={positive_semantics}
Negative/complement side: id={negative_id}; label={negative_label}; semantics={negative_semantics}

Allowed answer strings (canonical JSON; copy one item verbatim):
{allowed_answers}

Outcome distribution:
{outcome_distribution}

The prefilled JSON skeleton below contains every required canonical outcome_id exactly once. Return that complete skeleton; never return an empty mappings array and never delete, omit, duplicate, reorder, rename, or invent an object, outcome_id, or field. Change only each empty answer string. Every answer must be copied character-for-character from Allowed answer strings. Never use the generic substitutes positive, negative, true, or false; never add an explanation, qualification, paraphrase, or rewritten label. Do not change probabilities.

Return this prefilled JSON skeleton with only the answer string values changed:
{
 "mappings": [
   {
    "outcome_id": "{outcome_id}",
    "answer": ""
   }
 ]
}

Return exactly one valid JSON object. Do not use markdown fences, comments, trailing text, NaN, or Infinity.
```

Figure 10: Outcome-to-Answer Mapping Prompt.

B.2 Reasoning Direction Generation Prompt

This prompt is invoked once per query to record an LLM-side semantic prior with question-conditioned candidate entities and hypotheses. Candidate targets are selected from the admitted graph

entities, and the same call assigns each retained target to the positive or negative side of the canonical binary axis; targets without an unambiguous axis assignment are excluded. Causal distances are then computed independently by reverse BFS. The prompt is shown in Figure 8.

B.3 Multi-Outcome Probabilistic Reasoning Prompt

This prompt implements the LLM-side estimator f_{LLM}. Each round consumes q and graph context weighted by the prediction cutoff (entity narratives, contradictions, lightweight prompt-context causal chains, and previous-round outcomes), and emits a probability distribution over mutually exclusive resolution outcomes; the distribution is subsequently projected onto the binary probability P_{llm} by the outcome-to-answer mapping step. These prompt-context chains precede P_{llm} and are distinct from the subsequent CTVF-scored pool Π_q and its deduplicated subsets used to compute $P_{\mathrm{causal}}(q)$. The prompt is shown in Figure 9.

B.4 Outcome-to-Answer Mapping Prompt

This prompt finalizes the LLM-side estimate by aggregating the probability mass consistent with the binary resolution of the question, without modifying the underlying probabilities. The resulting scalar $P_{\mathrm{llm}} \in [0, 1]$ is subsequently combined with $P_{\mathrm{causal}}(q)$ via the adaptive fusion weight $\alpha(q)$. The prompt is shown in Figure 10.

C CHAIN Algorithm Details

Algorithm 1 presents the core calibration pipeline of **CHAIN**. The three phases of the algorithm map one-to-one to the three calibration bias sources targeted by our method, namely evidence weighting, chain aggregation, and source combination. **(1) Causal-Temporal Hypergraph Construction** builds the corpus-level hypergraph $\mathcal{G} = (V, E_H, \mathcal{C})$ once. For every factual chunk x, one extraction request to the extractor ϕ_{LLM} returns proposition/entity content and typed directed causal edges (with retries only on failure); the graph builder attaches the chunk-derived timestamp and optional stored similarity before populating $\mathcal{G}$ that all queries subsequently share. **(2) Per-Query Causal Estimation** addresses evidence-weighting bias via the CTVF and chain-aggregation bias via Direction-aware Noisy-OR. For each query, a cutoff-safe active view $\mathcal{G}_q$ retains known pre-cutoff evidence and evidence with missing or unparseable event time only when a verified availability upper bound precedes t_c. Admitted occurrences sharing the same cause, effect, and relation type are aggregated into a logical causal edge whose strongest occurrence supplies its weight. The reasoning-direction call receives the cutoff, canonical binary axis, and a bounded set of question-relevant entity labels from $\mathcal{G}_q$; keyword-filtered dense reranking then resolves $\mathcal{T}(q)$ with positive/negative axis-side orientations a_q, providing the BFS roots for the causal topological distance $d(\cdot, q)$. The distance drives the proximity exponent ρ, the temporal recency V_{rec}, the recurrence-salience V_{sal}, and the edge weight $\omega(c)$. A lightweight prompt context $\mathcal{H}_{\mathrm{ctx}}(q)$, comprising entity narratives, overall trajectory, contradictions, lightweight causal-context chains, source events, and prior-round outcomes, is supplied to f_{LLM} to obtain P_{llm}. These context chains are distinct from the formal CTVF chain pool Π_q constructed below. Forward breadth-first search enumerates node-simple paths from graph entities along outgoing causal edges; only paths that terminate at an oriented target are retained, and target nodes are not expanded. Chains are scored by $\Phi(\pi, q)$, partitioned by question-conditioned polarity $\delta_q(\pi)$, deduplicated by Jaccard-θ within each polarity, and merged through the neutral-anchored log-odds rule into $P_{\mathrm{causal}}(q)$. **(3) Adaptive Source Fusion** addresses source-combination bias. The unique-chain count k, the mean confidence $\bar{\Phi}(q)$, the saturated reliability $\kappa(k)$, and the floored balance $\widetilde{\beta}(q)$, computed on the deduplicated chains, are passed through the coverage map $\Omega(q)$ and a sigmoid gate to produce the question-specific weight $\alpha(q)$, which combines $P_{\mathrm{causal}}(q)$ with the LLM-side estimate $P_{\mathrm{llm}} = f_{\mathrm{LLM}}(q, \mathcal{H}_{\mathrm{ctx}}(q))$ into the final forecast $\hat{p}$.

Algorithm 1 CHAIN: Causal-Temporal Hypergraph Construction and Per-Query Calibration

Require: corpus of factual chunks $\{x\}$; query q with prediction cutoff t_c and canonical binary axis $\mathcal{A}_q$; extractor ϕ_{LLM}; estimator f_{LLM}; encoder $\mathrm{Enc}(\cdot)$; hyperparameters $d_{\max}, B_\pi, F_{\max}, \tau_\Phi, \theta, \eta, k_{\mathrm{sat}}, \beta_0, Z, \alpha_b, \alpha_0, \Omega_0, \zeta$; fixed v_{fb} and $\varepsilon = 10^{-6}$

Ensure: forecast $\hat{p} \in [0, 1]$

// 1: Causal-Temporal Hypergraph Construction

Initialize $V \leftarrow \emptyset$, $E_H \leftarrow \emptyset$, $\mathcal{C} \leftarrow \emptyset$

for each chunk x **do**

  $(\mathcal{P}_x, \mathcal{C}_x) \leftarrow \phi_{\mathrm{LLM}}(x)$

  Attach to each $e \in \mathcal{P}_x$ the timestamp parsed from x and optional stored similarity s_e

  $E_H \leftarrow E_H \cup \mathcal{P}_x$; $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_x$; $V \leftarrow V \cup \bigcup_{(V_e,t,s)\in\mathcal{P}_x} V_e$

end for

Encode $\mathrm{Enc}(v)$ for $v \in V$ and $\mathrm{Enc}(e)$ for $e \in E_H$

// 2: Per-Query Causal Estimation

$\mathcal{G}_q = (V_q, E_{H,q}, \mathcal{C}_q) \leftarrow \mathrm{CutoffSafeActiveView}(\mathcal{G}, q, t_c)$ using verified pre-cutoff admissibility and typed-edge aggregation

$\mathcal{L}_q \leftarrow$ bounded question-relevant entity labels from $\mathcal{G}_q$

$\mathcal{D}(q) \leftarrow \mathrm{ReasoningDirection}(q, t_c, \mathcal{A}_q, \mathcal{L}_q)$

$(\mathcal{T}(q), a_q) \leftarrow \mathrm{ResolveAnchors}(\mathcal{D}(q), \mathcal{G}_q)$ by keyword-filtered dense reranking; discard weak or conflicting matches

$d(v, q) \leftarrow$ exact reverse-BFS distance from $\mathcal{T}(q)$ on $\mathcal{C}_q$ within depth $d_{\max}$ for $v \in V_q$; set $d(v, q) = \infty$ otherwise

$d(e, q) \leftarrow \min_{v\in V_e} d(v, q)$ for each $e = (V_e, t, s) \in E_{H,q}$

$\rho(n, q) \leftarrow \min(d(n, q)/d_{\max},\ 1)$ for $n \in V_q \cup E_{H,q}$

for each $e = (V_e, t, s) \in E_{H,q}$ **do**

  Set $N(e)$ to the number of distinct admitted records sharing its canonicalized proposition key

  if t is parseable and $t < t_c$ **then**

    $V_{\mathrm{rec}}(e, q) \leftarrow \exp\big(-\frac{1}{2}\,\rho(e, q)\,\ln\max\{1, (t_c - t)_{\mathrm{days}}\}\big)$; $V_{\mathrm{sal}}(e) \leftarrow \mu(N(e))\,\sqrt{N(e)}\,M_r(e)$

    $V_{\mathrm{CTVF}}(e, q) \leftarrow \min(V_{\mathrm{rec}}(e, q) + V_{\mathrm{sal}}(e),\ 1)$

  else

    $V_{\mathrm{CTVF}}(e, q) \leftarrow v_{\mathrm{fb}}$ (only with a verified availability bound before t_c)

  end if

end for

$V_{\mathrm{CTVF}}(v, q) \leftarrow$ mean of fixed incident-hyperedge scores $V_{\mathrm{CTVF}}(e, q)$ for $v \in V_q$ (else neutral 0.5)

$\mathcal{H}_{\mathrm{ctx}}(q) \leftarrow \mathrm{PromptContext}(\mathcal{G}_q, q, t_c)$; $P_{\mathrm{llm}} \leftarrow f_{\mathrm{LLM}}(q, \mathcal{H}_{\mathrm{ctx}}(q))$

for each $c = (u, v, \mathrm{type}, w) \in \mathcal{C}_q$ **do**

  $\rho_c \leftarrow \min\big((d(u, q) + d(v, q))/(2\,d_{\max}),\ 1\big)$; $\omega(c) \leftarrow \big(\frac{1}{2}(V_{\mathrm{CTVF}}(u, q) + V_{\mathrm{CTVF}}(v, q))\big)^{\rho_c}$

end for

Forward breadth-first expand node-simple paths from each entity to depth $d_{\max}$ using at most the $F_{\max}$ strongest outgoing edges; set $\widetilde{\Phi}(\pi, q) \leftarrow \prod_\ell w_\ell \omega(c_\ell)$ and prune when $\widetilde{\Phi}(\pi, q) < \tau_\Phi$

Stop enumeration after $H_\pi \leftarrow \max(5B_\pi, 1000)$ score-qualified target-terminating paths have been collected

Let Π_q be the top B_π target-terminating paths and split by $\delta_q(\pi) = a_q(v_L)(-1)^{P(\pi)}$ into Π_q^+, Π_q^-

Greedy Jaccard-θ deduplication within each polarity $\rightarrow \widehat{\Pi}_q^+$, $\widehat{\Pi}_q^-$

$p^+(q) \leftarrow 1 - \prod_{\pi\in\widehat{\Pi}_q^+}(1 - \Phi(\pi, q))$; $p^-(q) \leftarrow 1 - \prod_{\pi\in\widehat{\Pi}_q^-}(1 - \Phi(\pi, q))$

$P_{\mathrm{causal}}(q) \leftarrow \sigma\big(g_\varepsilon(p^+(q)) - g_\varepsilon(p^-(q))\big)$, where $g_\varepsilon(p) = \mathrm{logit}(\min(1 - \varepsilon, \max(\varepsilon, (1 + p)/2)))$

// 3: Adaptive Source Fusion

$k \leftarrow |\widehat{\Pi}_q^+| + |\widehat{\Pi}_q^-|$; $\bar{\Phi}(q) \leftarrow \frac{1}{\max(k,1)} \sum_{\pi\in\widehat{\Pi}_q^+\cup\widehat{\Pi}_q^-} \Phi(\pi, q)$

$\kappa(k) \leftarrow \min(\ln(1 + k)/\ln(1 + k_{\mathrm{sat}}),\ 1)$

$\beta(q) \leftarrow \min(|\widehat{\Pi}_q^+|, |\widehat{\Pi}_q^-|)/\max(|\widehat{\Pi}_q^+|, |\widehat{\Pi}_q^-|, 1)$; $\widetilde{\beta}(q) \leftarrow \beta_0 + (1 - \beta_0)\,\beta(q)$

$\Omega(q) \leftarrow \min(k\,\bar{\Phi}(q)\,\kappa(k)\,\widetilde{\beta}(q)/Z,\ 1)$

$\alpha(q) \leftarrow \mathbb{1}[k > 0] \cdot \min(2\alpha_b\,\sigma(\zeta\,(\Omega(q) - \Omega_0)),\ \alpha_0)$

return $\hat{p} = \alpha(q)\,P_{\mathrm{causal}}(q) + (1 - \alpha(q))\,P_{\mathrm{llm}}$

Computational Complexity. The computational cost of **CHAIN** can be decomposed into three stages. *Hypergraph construction* issues one extraction request per factual chunk, with retries only on failure, and scales linearly with the corpus size; this cost is incurred only once per corpus and is amortized over all downstream queries. Per-query construction of $\mathcal{G}_q$ adds one graph scan plus deterministic grouping of admitted occurrences. *Per-query causal estimation* performs a reverse

breadth-first search over $\mathcal{C}_q$ for each target in $\mathcal{T}(q)$, which runs in $\mathcal{O}(|\mathcal{T}(q)|(|V_q|+|\mathcal{C}_q|))$ in the worst case and is bounded by the depth $d_{\max}$, followed by breadth-first chain enumeration with uncapped worst-case complexity $\mathcal{O}(|V_q|\, b^{d_{\max}})$ for average fan-out b. The controls $F_{\max}$ and τ_Φ restrict expansion, $H_\pi = \max(5B_\pi, 1000)$ caps score-qualified enumeration, B_π bounds $|\Pi_q|$, and Jaccard-θ deduplication determines k. Thus, the polarity-wise Noisy-OR aggregation and the subsequent neutral-anchored combination admit $\mathcal{O}(k)$ complexity. *Adaptive fusion* runs in $\mathcal{O}(1)$ given the chain statistics k, $\bar{\Phi}(q)$, and the polarity counts. In practice, the per-query runtime is dominated by the language-model invocations rather than by graph traversal, encompassing the reasoning-direction generation, the multi-outcome estimation performed by f_{LLM}, and the outcome-to-answer mapping.

D Dataset Details

We conduct experiments on eight public forecasting datasets, covering both in-distribution and out-of-distribution sources, and spanning forecasting tasks across diverse domains.

AI-Futures (Turtel et al., 2026; LightningRodLabs, 2025a): A forecasting dataset of AI-domain questions on AI labs, model releases, agentic deployments, and AI-industry developments, generated under the future-as-label paradigm so that ground-truth labels arise from real-world events post-dating the model knowledge cutoff.

Metaculus (Chandak et al., 2026): A forecasting dataset of community-resolved questions extracted from the Metaculus forecasting platform, with explicit closing and resolution times and forecaster-count metadata.

Polymarket (Polymarket, 2024): A market-driven forecasting dataset of prediction-market questions sourced from Polymarket, where each question is backed by a tradable contract whose outcome is settled by a real-world event.

Future-as-Label (Turtel et al., 2026): A forecasting dataset of mixed-domain questions whose ground-truth labels are determined by future real-world outcomes occurring after the model knowledge cutoff, ensuring that the task cannot be solved by memorization.

Clinical Trial Outcomes (Gao et al., 2024): A large-scale benchmark of clinical trial outcomes for drug development, providing trial-level success/failure labels and introducing a biomedical out-of-distribution shift relative to the in-distribution forecasting sources.

FOReCAst (Yuan et al., 2025): A future-outcome reasoning and confidence-assessment benchmark spanning Boolean, timeframe, and quantity question types; we use its Boolean subset, in which each question requires the model to predict an outcome together with a confidence estimate.

Golf-Forecasting (Turtel et al., 2026; LightningRodLabs, 2025b): A forecasting dataset of questions about professional golf tournaments, generated under the future-as-label paradigm with news articles as the underlying source, providing a fast-resolving sports out-of-distribution setting.

KalshiBench (Nel, 2025): A forecasting benchmark sourced from the Kalshi regulated event-contract exchange, covering sports, politics, macroeconomics, and other event categories, designed to evaluate epistemic calibration on prediction-market questions.

We designate the first four datasets as in-distribution sources and the remaining four as held-out-source OOD datasets spanning diverse forecasting domains, and choose three backbone LLMs whose knowledge cutoffs predate the resolution dates of nearly all sampled questions, so that calibration performance reflects forward-only forecasting rather than memorization. FOReCAst contains some potentially overlapping cases whose resolution dates precede the model knowledge cutoffs, but this leakage risk applies uniformly to all baselines and does not favor **CHAIN**.

To balance the cost of hypergraph construction against evaluation coverage, we construct a separate causal-temporal hypergraph for each of the four in-distribution datasets from 2,048 instances sampled from that dataset. We also randomly sample 128 test instances from each of the eight datasets; these test instances are disjoint from the instances used for hypergraph construction. For the main OOD evaluation reported in Table 2, the four out-of-distribution datasets do not use dataset-specific retrieval corpora. Instead, we take the union of the source documents of the four in-distribution datasets as a single unified ID corpus, and each retrieval-based method (NaiveRAG, GraphRAG, LightRAG, HippoRAG, HyperGraphRAG, and **CHAIN**) builds its own retrieval structure on top of this unified

corpus using its own native construction pipeline. This setup keeps the upstream data identical across all methods while letting each method retain its own structural representation, so observed OOD calibration differences reflect algorithmic rather than corpus differences. Although the OOD datasets are held out at the data-source level, their topical domains can overlap with those of the in-distribution datasets. This setup therefore evaluates cross-source generalisation under a unified ID retrieval corpus rather than strictly domain-disjoint transfer; the degree of topical mismatch varies by target dataset.

E Baseline Details

We compare CHAIN with 13 baselines spanning direct probability elicitation, step-by-step reasoning, retrieval-augmented generation, and post-hoc probability calibration. The following descriptions summarize each method's input protocol, structural representation, or calibration mapping.

LLM Direct (Ouyang et al., 2022) prompts the LLM with the raw question and asks it to output a probability and answer in a single shot.

LLM CoT (Wei et al., 2022) prepends a chain-of-thought prompt so that the LLM produces step-by-step reasoning before its final probability and answer.

AutoCast-style Prompting (Zou et al., 2022) uses the forecasting setup of AutoCast, providing the LLM with the question together with associated background news context as supporting information.

NaiveRAG (Lewis et al., 2020) follows the standard retrieval-augmented generation paradigm, performing chunk-level dense retrieval over the corpus and providing the top-k chunks to the LLM as supporting context.

GraphRAG (Edge et al., 2024) extracts an entity-level knowledge graph from the corpus, applies hierarchical community detection, and retrieves pre-generated community summaries to support query-focused generation.

LightRAG (Guo et al., 2025) builds a compact entity-relation graph and performs dual-level retrieval that combines a low-level entity-specific view and a high-level theme-level view of the graph.

HippoRAG (Gutiérrez et al., 2024) adopts a neurobiologically inspired memory index, applying personalized PageRank over an OpenIE-style entity-and-triple graph to surface multi-hop evidence.

HyperGraphRAG (Luo et al., 2025) represents knowledge as n-ary relational hyperedges and performs hypergraph-structured retrieval that fuses entity-level and hyperedge-level evidence.

Temperature Scaling (Guo et al., 2017) learns a single scalar temperature on a validation set to rescale logits before the softmax.

Platt Scaling (Platt, 1999) fits a sigmoid mapping from uncalibrated scores to class probabilities via maximum likelihood.

Isotonic Regression (Zadrozny & Elkan, 2002) fits a non-parametric monotone mapping from predicted confidences to empirical class probabilities.

Histogram Binning (Zadrozny & Elkan, 2001) partitions the confidence axis into bins and replaces each bin's prediction with the empirical class frequency observed in that bin on the validation set.

Conformal-Score Probability Adjustment (Angelopoulos & Bates, 2023) is a conformal-inspired scalar baseline. From the calibration predictions, we compute nonconformity scores $r_i = |p_i - y_i|$ and their 0.9 quantile q; each evaluation probability is then mapped to $\widetilde{p} = 0.5 + (1 - q)(p - 0.5)$. It outputs adjusted scalar probabilities rather than prediction sets and makes no conformal coverage claim.

F Evaluation Metrics

We evaluate calibration quality and predictive accuracy across seven metrics spanning complementary dimensions of probabilistic forecasting. ECE, ACE, MCE, and Reliability quantify calibration error; NLL measures the logarithmic loss assigned to observed outcomes; Brier score is a proper scoring rule capturing both calibration and sharpness; and Accuracy measures the correctness of

binary predictions. Let $\{(p_i, y_i)\}_{i=1}^N$ denote N predictions, where $p_i \in [0,1]$ is the predicted probability of the positive event and $y_i \in \{0,1\}$ is the ground-truth label. For each binning scheme, $B_m \subseteq \{1,\dots,N\}$ denotes the indices assigned to bin m, with $M = 10$, mean predicted event probability $\bar{p}_m = |B_m|^{-1} \sum_{i \in B_m} p_i$, and empirical event frequency $\bar{y}_m = |B_m|^{-1} \sum_{i \in B_m} y_i$. Empty bins are omitted from all bin-based sums and maxima.

ECE (Expected Calibration Error) (Guo et al., 2017). We compute the bin-weighted gap between the mean predicted event probability and empirical event frequency under equal-width binning of $[0,1]$:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \big|\bar{p}_m - \bar{y}_m\big|. \tag{78}$$

Lower values indicate better global calibration.

ACE (Adaptive Calibration Error) (Nixon et al., 2019). We use an adaptive-binning calibration error implemented with approximately equal-mass bins:

$$\mathrm{ACE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \big|\bar{p}_m - \bar{y}_m\big|. \tag{79}$$

Lower values indicate better calibration when predictions are unevenly distributed across the predicted-probability axis.

MCE (Maximum Calibration Error) (Guo et al., 2017). Using the same equal-width bins as ECE, we take the worst bin-wise calibration gap, restricting attention to bins containing at least five samples for stability:

$$\mathrm{MCE} = \max_{m:\,|B_m| \geq 5} \big|\bar{p}_m - \bar{y}_m\big|. \tag{80}$$

Lower values indicate that no single predicted-probability region exhibits severely miscalibrated predictions.

Rel (Reliability) (Murphy, 1973). Using the same equal-width bins as ECE, we report the binned estimate of the reliability component of the Brier-score decomposition:

$$\mathrm{Rel} = \sum_{m=1}^{M} \frac{|B_m|}{N} \big(\bar{p}_m - \bar{y}_m\big)^2. \tag{81}$$

Lower values indicate that bin-wise mean event probabilities are closer to the corresponding empirical frequencies.

NLL (Negative Log-Likelihood). We compute the mean binary logarithmic loss assigned to the observed outcomes:

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \widetilde{p}_i + (1 - y_i) \log(1 - \widetilde{p}_i) \right], \tag{82}$$

where $\widetilde{p}_i = \min(\max(p_i, \epsilon), 1 - \epsilon)$ and $\epsilon = 10^{-7}$ prevents numerical singularities. Lower values indicate better probabilistic predictions.

Brier (Brier Score) (Brier, 1950). We compute the mean squared deviation between the predicted probability and the binary outcome:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2. \tag{83}$$

Lower values indicate predictions that are jointly accurate and well-calibrated.

Acc (Accuracy). We report the fraction of correctly predicted labels under a 0.5 decision threshold:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i], \qquad \hat{y}_i = \mathbb{1}[p_i \geq 0.5]. \tag{84}$$

Higher values indicate more accurate binary predictions.

ECE, ACE, MCE, Rel, and Brier are reported in percent and are lower-better; Accuracy is reported in percent and is higher-better. NLL is reported on its dimensionless natural-log scale and is lower-better.

Table 4: Comparison of parameter settings across the three baseline categories and **CHAIN**.

Group	Parameter	Direct	RAG	Post-hoc	CHAIN
Models	Backbone LLMs	3 LLMs	3 LLMs	3 LLMs	3 LLMs
	Construction LLM	–	GPT-4o-mini	–	GPT-4o-mini
	Encoder	–	text-embedding-3-small	–	text-embedding-3-small
	Embedding dim	–	1536	–	1536
Inference	Temperature	0.3	0.3	0.3	0.3
	Max tokens	8192	8192	8192	8192
	Inference API	official	official	official	official
Retrieval	Chunk size	–	512 tokens	–	512 tokens
	Chunk overlap	–	64 tokens	–	64 tokens
	RAG top-k	–	10	–	–
	Similarity	–	cosine	–	cosine

Table 5: Core **CHAIN**-specific hyperparameters and default values used unless otherwise specified.

Symbol	Meaning	Value	Sensitivity
$d_{\max}$	reachability/search depth cap	4	swept $\{2, 3, 4, 5\}$
θ	Jaccard threshold for same-polarity dedup	0.5	swept $\{0.3, 0.5, 0.7\}$
η	recurrence saturation temperature in V_{sal}	5.0	—
$M_r(e)$	default when stored similarity is unavailable	0.5	—
v_{fb}	cutoff-safe missing-time CTVF fallback	0.25	—
k_{sat}	reliability saturation anchor	10	—
β_0	balance floor	0.3	—
Z	coverage normalizer	5.0	—
α_b	mapping scale	0.5	—
α_0	fusion upper bound $(1.2\alpha_b)$	0.6	—
Ω_0	coverage activation threshold	0.3	—
ζ	sigmoid sharpness	3.0	—

G IMPLEMENTATION DETAILS

We construct the causal-temporal hypergraph using GPT-4o-mini (Hurst et al., 2024) as the extractor and `text-embedding-3-small` as the shared encoder, and evaluate **CHAIN** on three LLM backbones: DeepSeek-V3 (DeepSeek-AI et al., 2025), GPT-4o-mini (Hurst et al., 2024), Gemini-2.0-Flash (Kavukcuoglu, 2025). All experiments are conducted via the official APIs. For **CHAIN** and all baselines, the reported comparison results are averaged over three independent runs. To ensure a controlled comparison, every method is run under the unified experimental setup summarized in Table 4, including matched decoding parameters, the same evaluation set for each dataset, and the same backbone LLMs. For all baselines, we adopt the prompts from their official implementations. The **CHAIN**-specific hyperparameters that govern causal proximity, deduplication, evidence aggregation, and source fusion are listed in Table 5. We use the listed defaults across all eight datasets and three backbone LLMs unless otherwise specified in the sensitivity analyses.

Prompt Matching and Baseline Fairness. To ensure comparability across forecasting methods, Direct, CoT, NaiveRAG, GraphRAG, HyperGraphRAG, and **CHAIN** follow a unified input and evaluation protocol. All methods receive the same question statement, prediction cutoff, and admissible corpus boundary. For questions with multiple candidate outcomes, we apply the same outcome-decomposition rule to map the task into a common binary-event prediction space. We also use consistent option representations, probability-output formats, and result-parsing rules across methods, thereby controlling differences in prompt structure, output constraints, and post-processing. For retrieval-augmented methods, the same cutoff and corpus boundary apply, and all records directly associated with test questions are excluded from the retrieval corpus. The resulting evaluation compares the reasoning, retrieval structure, causal-temporal hypergraph construction, and evidence aggregation used by each method under the same protocol.

Although the hypergraph is extracted with GPT-4o-mini for all datasets, **CHAIN**'s gain over the strongest baseline transfers stably across backbone LLMs. For each backbone and each evaluation metric, we identify the strongest baseline as the one with the best mean across the four in-distribution datasets, and report the gap between **CHAIN**'s mean and this baseline's mean. Across GPT-4o-mini, Gemini-2.0-Flash, and DeepSeek-V3, respectively, **CHAIN** reduces ECE by 4.77, 4.15, and 3.07 points, reduces Brier score by 5.57, 4.34, and 3.30 points, and improves Accuracy by 9.57, 8.20, and

Table 6: **CHAIN**'s gain over the strongest baseline by backbone LLM. For each backbone and each metric, the strongest baseline is identified by its mean over the four in-distribution datasets, and the reported value is the gap between **CHAIN**'s mean and this baseline's mean. The hypergraph is extracted with GPT-4o-mini in all rows.

Backbone	ΔECE ↓	ΔBrier ↓	ΔAcc ↑
GPT-4o-mini	$\mathbf{-4.77}$	$\mathbf{-5.57}$	$\mathbf{+9.57}$
Gemini-2.0-Flash	-4.15	-4.34	$+8.20$
DeepSeek-V3	-3.07	-3.30	$+7.42$

Table 7: Performance comparison of **CHAIN** and the baselines on 4 OOD datasets with DeepSeek-V3.

Method	CTO			FOReCAst			Golf-Forecasting			KalshiBench			Average		
	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑	ECE↓	Brier↓	Acc↑
LLM Direct	15.47	25.80	60.16	9.72	20.03	72.66	14.77	26.38	57.03	16.09	25.30	60.94	14.01	24.38	62.70
LLM CoT	16.52	25.01	57.81	10.80	20.28	71.09	17.76	29.48	54.69	20.16	25.54	58.59	16.31	25.08	60.55
AutoCast	12.50	24.27	61.72	16.33	25.61	57.81	13.32	27.53	56.25	13.24	26.28	53.91	13.85	25.92	57.42
NaiveRAG	19.84	25.79	60.16	19.56	27.31	60.16	11.13	26.63	49.22	15.59	28.68	46.88	16.53	27.10	54.10
GraphRAG	13.28	25.29	59.38	8.55	20.77	71.09	14.22	26.61	53.91	18.74	26.46	53.12	13.70	24.78	59.38
LightRAG	17.15	26.16	53.12	11.79	21.58	71.88	12.77	26.98	51.56	20.39	25.48	56.25	15.53	25.05	58.20
HippoRAG	20.04	27.48	50.00	13.23	21.04	71.09	20.47	28.17	56.25	28.40	30.99	46.09	20.54	26.92	55.86
HyperGraphRAG	21.48	26.54	50.00	13.24	21.04	71.09	13.63	28.56	48.44	23.95	26.26	53.12	18.08	25.60	55.66
TempScaling	4.05	23.71	60.16	9.72	20.19	72.66	5.23	24.35	57.03	7.52	23.04	60.94	6.63	22.82	62.70
PlattScaling	30.08	32.95	42.19	27.97	29.96	46.88	32.01	34.34	44.53	34.93	35.11	39.84	31.25	33.09	43.36
IsotonicReg	28.57	31.87	48.44	25.02	26.47	67.19	29.91	33.01	50.78	30.69	32.26	57.03	28.55	30.90	55.86
HistogramBin	36.45	44.86	44.53	27.30	32.08	62.50	31.03	35.59	57.81	33.19	38.78	53.91	31.99	37.83	54.69
ConformalAdj	3.91	23.97	60.16	15.34	22.02	72.66	5.16	24.35	57.03	8.12	23.58	60.94	8.13	23.48	62.70
CHAIN (ours)	**3.61**	**22.73**	**62.50**	**6.98**	**17.90**	**74.22**	**5.15**	**23.91**	**58.59**	**7.25**	**22.95**	**62.50**	**5.75**	**21.87**	**64.45**

7.42 points; per-backbone gains are reported in Table 6. The evaluated search-control configurations and the defaults adopted by **CHAIN** are reported in Table 8.

H Results of DeepSeek-V3

Table 7 reports the complete results of DeepSeek-V3 on Clinical Trial Outcomes, FOReCAst, Golf-Forecasting, and KalshiBench under the same unified in-distribution retrieval-corpus protocol used for the main OOD evaluation. **CHAIN** achieves the lowest ECE and Brier score and the highest Accuracy on each dataset and in the four-dataset macro average. On the macro average, **CHAIN** reduces ECE from 6.63 to 5.75 and Brier score from 22.82 to 21.87, while increasing Accuracy from 62.70 to 64.45, relative to Temperature Scaling, the strongest post-hoc baseline in this comparison. These results show that the calibration and predictive-performance gains persist with DeepSeek-V3 across all four OOD datasets.

Table 8: Evaluated search-budget, depth, fan-out, and pruning-threshold settings, including the defaults used in the main experiments.

Parameter	Evaluated settings	Default
Chain budget B_π	50, 100, 200, 400	200
Maximum search depth $d_{\max}$	2, 3, 4, 5	4
Fan-out cap $F_{\max}$	2, 4, 8, 16, unlimited	unlimited
Pruning threshold τ_Φ	0, 0.001, 0.005, 0.02, 0.05	0.005

I Search Configuration and Graph-Scale Sensitivity

I.1 Search Configuration

We examine the effects of the chain budget B_π, maximum search depth $d_{\max}$, per-node outgoing-edge cap $F_{\max}$, and low-score prefix pruning threshold τ_Φ. We use DeepSeek-V3 and report macro-averaged results over AI-Futures, Future-as-Label, Clinical Trial Outcomes, and Golf-Forecasting. All other settings remain fixed at their defaults. Table 8 summarizes the evaluated settings and the

Table 9: Predictive performance under different search configurations.

Parameter	Setting	ECE ↓	Brier ↓	Accuracy ↑
Chain budget B_π	50	6.43	20.13	67.58
	100	6.62	20.67	67.58
	$200^\dagger$	6.85	20.70	67.58
	400	6.68	20.70	67.58
Maximum search depth $d_{\max}$	2	6.53	20.14	67.97
	3	5.90	20.47	67.58
	$4^\dagger$	6.85	20.70	67.58
	5	6.60	20.81	67.38
Outgoing-edge cap $F_{\max}$	2	7.04	20.37	68.16
	4	6.32	20.23	68.36
	8	6.01	20.40	67.77
	16	6.49	20.21	67.97
	Unlimited†	6.85	20.70	67.58
Pruning threshold τ_Φ	0	6.85	20.70	67.58
	0.001	6.85	20.70	67.58
	$0.005^\dagger$	6.85	20.70	67.58
	0.02	6.83	20.70	67.58
	0.05	6.88	20.72	67.77

Table 10: Candidate-chain counts and computational cost under different pruning thresholds. Relative runtime is normalized by $\tau_\Phi = 0$.

τ_Φ	Pre-dedup.	Post-dedup.	Mean time (s)	Relative time ↓
0	30.76	13.31	10.48	1.00
$0.005^\dagger$	30.76	13.31	2.47	0.24
0.02	29.68	13.15	1.70	0.16
0.05	24.16	12.32	1.25	0.12

defaults used in the main experiments. Table 9 reports the corresponding predictive performance, where † denotes the default adopted by **CHAIN**.

Accuracy remains 67.58 across all chain budgets, while ECE and Brier vary only slightly, showing stable predictions across changes in the candidate-chain pool size. Performance also remains similar as the maximum search depth and outgoing-edge cap change. Some smaller depth and edge-cap settings improve all three metrics relative to their defaults, but the best ECE, Brier, and Accuracy are attained by different configurations. Thus, **CHAIN** maintains stable predictive quality across different search spaces without depending on a single search configuration.

The pruning results show similar stability. Disabling pruning yields an ECE of 6.85, a Brier score of 20.70, and an Accuracy of 67.58, identical to the default threshold $\tau_\Phi = 0.005$. Across the evaluated thresholds, ECE is 6.83–6.88, Brier is 20.70–20.72, and Accuracy is 67.58–67.77. Changing the threshold therefore does not cause a material loss in predictive performance.

Table 10 further reports the numbers of candidate chains before and after Jaccard deduplication, together with the mean runtime of offline path search and prediction aggregation. Relative runtime is normalized by the no-pruning setting $\tau_\Phi = 0$. These measurements use fixed graphs and fixed model outputs and exclude graph construction and language-model calls.

The default threshold $\tau_\Phi = 0.005$ produces the same predictions as disabling pruning while reducing the mean search-and-aggregation time from 10.48 to 2.47 seconds. Increasing the threshold to 0.02 and 0.05 further reduces both candidate-chain counts and runtime while preserving predictive quality. Low-score prefix pruning terminates subsequent expansion of weak paths, reducing path-enumeration and aggregation costs while retaining the principal causal-temporal evidence.

Overall, the predictive results remain stable as the chain budget, maximum search depth, outgoing-edge cap, and pruning threshold change. Low-score prefix pruning additionally reduces the cost of path search and prediction aggregation.

Table 11: Predictive performance under different causal-temporal graph scales. Graph scale denotes the configured source-record sampling fraction; results are macro-averaged over four datasets.

Source-record fraction	ECE ↓	Brier ↓	Accuracy ↑
25%	8.62	22.24	63.87
50%	9.24	21.87	64.45
75%	8.86	21.58	65.23
100%	**6.85**	**20.70**	**67.58**

I.2 Graph Scale

We next examine how the coverage of the causal-temporal graph affects predictive performance. For each corpus source, we use a fixed random order to construct nested 25%, 50%, 75%, and 100% source-record subsets. Each larger subset contains the records retained by the smaller subsets, together with all text chunks, nodes, and causal relations associated with the selected records. Subset construction does not use prediction labels or model outputs.

As the source-record fraction increases from 25% to 100%, the Brier score decreases monotonically from 22.24 to 20.70, while Accuracy increases monotonically from 63.87 to 67.58. ECE is not strictly monotonic at the intermediate scales, but the complete graph still achieves the lowest ECE and the best Brier score and Accuracy simultaneously.

The graph variants are nested, label-independent source-record subsets, with the forecasting model, inference procedure, and evaluation protocol held fixed. Expanding causal-temporal evidence coverage supplies **CHAIN** with more complete events, entities, and causal relations and improves probabilistic prediction quality. The complete graph achieves the best result on all three metrics, demonstrating that **CHAIN** can use the additional causal-temporal information to produce more accurate and better-calibrated predictions.

Together, the search results show stable predictive quality under different path-search configurations and lower search cost from low-score prefix pruning, while the graph-scale results show improved performance with broader causal-temporal evidence coverage. **CHAIN** therefore controls path-search cost while effectively using a more complete evidence graph to improve predictive quality.

J Robustness to Hypergraph Extractor Choice

To assess the robustness of **CHAIN** to the choice of hypergraph extractor, we separately construct the causal-temporal hypergraph with GPT-4o-mini and GPT-4.1-nano while keeping the DeepSeek-V3 forecasting backbone, input data, inference pipeline, and evaluation protocol fixed. This evaluation covers all eight forecasting datasets: AI-Futures, Metaculus, Polymarket, Future-as-Label, Clinical Trial Outcomes, FOReCAst, Golf-Forecasting, and KalshiBench. We denote the graph-free and retrieval-free forecasting control evaluated under the same protocol as Direct Multi-Outcome forecasting (Direct-MO). For each question, it decomposes the resolution space into mutually exclusive and collectively exhaustive outcomes, assigns a probability to each outcome, projects the resulting distribution onto the canonical binary axis, and sums the probability mass consistent with the target event to obtain the final event probability.

Per-dataset results under the two extractors are reported in Table 12, and the eight-dataset macro-average comparison is reported in Table 13. With GPT-4.1-nano extraction, **CHAIN** reduces macro-average ECE and Brier by 5.32 and 2.37 percentage points relative to Direct-MO and improves Accuracy by 7.23 percentage points, showing that its calibration and predictive advantages persist with a smaller hypergraph extractor.

With the default GPT-4o-mini extractor, macro-average ECE and Brier further decrease to 8.14 and 19.89, while Accuracy increases to 70.02%. Relative to GPT-4.1-nano, GPT-4o-mini lowers ECE and Brier by 1.75 and 1.21 percentage points and improves Accuracy by 2.54 percentage points.

CHAIN outperforms Direct-MO in macro-average ECE, Brier, and Accuracy under both extractors, while the default GPT-4o-mini extractor achieves the best result on all three metrics. Therefore, **CHAIN** maintains its calibration and predictive advantages across the evaluated hypergraph extractors and achieves higher overall predictive quality with its default extractor.

Table 12: Performance of **CHAIN** with different hypergraph extractors across eight datasets.

Dataset	Hypergraph extractor	ECE↓	Brier↓	Accuracy↑
AI-Futures	GPT-4o-mini	8.54	17.34	76.56
AI-Futures	GPT-4.1-nano	10.38	17.94	73.44
Metaculus	GPT-4o-mini	9.72	18.05	73.44
Metaculus	GPT-4.1-nano	12.26	19.40	74.22
Polymarket	GPT-4o-mini	13.74	17.41	79.69
Polymarket	GPT-4.1-nano	11.21	21.49	67.97
Future-as-Label	GPT-4o-mini	10.09	18.82	72.66
Future-as-Label	GPT-4.1-nano	6.11	20.38	68.75
Clinical Trial Outcomes	GPT-4o-mini	3.61	22.73	62.50
Clinical Trial Outcomes	GPT-4.1-nano	5.24	22.97	61.72
FOReCAst	GPT-4o-mini	6.98	17.90	74.22
FOReCAst	GPT-4.1-nano	11.62	18.96	71.88
Golf-Forecasting	GPT-4o-mini	5.15	23.91	58.59
Golf-Forecasting	GPT-4.1-nano	13.21	24.49	62.50
KalshiBench	GPT-4o-mini	7.25	22.95	62.50
KalshiBench	GPT-4.1-nano	9.07	23.14	59.38

Table 13: Macro-averaged performance across eight datasets under different hypergraph-extractor settings.

Method	Extractor	ECE↓	Brier↓	Accuracy↑
Direct-MO	None	15.21	23.47	60.25
CHAIN	GPT-4.1-nano	9.89	21.10	67.48
CHAIN	**GPT-4o-mini**	**8.14**	**19.89**	**70.02**

K COMPARISON WITH MODERN POST-HOC CALIBRATION BASELINES

To compare **CHAIN** with modern post-hoc calibration methods, we apply Temperature Scaling, Platt Scaling, Isotonic Regression, Beta Calibration, Venn–Abers, and Binary Dirichlet calibration to the corresponding LLM Direct predictions. All post-hoc methods use a leakage-safe fitting protocol. For in-distribution data, we use five-fold cross-fitting, so the calibrated prediction for each evaluated instance is produced by a calibrator fitted without that instance's label. For out-of-distribution data, each calibrator is fitted only on the pooled in-distribution data for the corresponding backbone and then applied directly to the out-of-distribution predictions. In contrast, **CHAIN** performs calibration at inference time using structured causal evidence. The table reports macro averages over the 24 dataset–backbone configurations formed by three backbones and eight datasets.

Because all outcomes in this evaluation are binary, our Binary Dirichlet implementation uses the same unconstrained log-probability feature model as Beta Calibration, with the same fitting and regularization protocol. The two methods are therefore mathematically equivalent under this binary setting, and their identical values are expected.

The results in Table 14 show that **CHAIN** achieves the lowest ECE, Brier score, and NLL and the highest Accuracy relative to all listed post-hoc calibration baselines. Temperature Scaling performs best among the listed post-hoc methods on all four metrics. Relative to Temperature Scaling, **CHAIN** lowers ECE and Brier score by 4.28 and 3.55 percentage points, respectively, lowers NLL by 0.0810, and improves Accuracy by 7.98 percentage points. The experimental results show that inference-time calibration based on structured causal evidence outperforms the evaluated post-hoc calibration baselines in both probability quality and predictive accuracy.

L CONTROLLING FOR LLM FORECASTING CALLS WITH MULTI-SAMPLE DIRECT-MO

To test whether increasing the number of LLM forecasts can attain the performance gains of **CHAIN**, we compare a single Direct-MO forecast, the probability average of five independent Direct-MO forecasts, and the complete **CHAIN** method on all eight datasets with DeepSeek-V3. Table 15 reports the average number of LLM calls and tokens per evaluation instance, together with macro-averaged ECE, Brier score, NLL, and Accuracy across the eight datasets. All methods use the same test set for

Table 14: Comparison with modern post-hoc calibration methods under a leakage-safe fitting protocol.

Method	ECE↓	Brier↓	NLL↓	Acc↑
Temperature Scaling	12.81	23.72	0.6714	61.49
Platt Scaling	20.21	26.69	0.7405	54.43
Isotonic Regression	20.27	26.73	0.7860	52.90
Beta Calibration	20.61	26.64	0.7389	53.12
Venn–Abers	20.04	26.60	0.7384	52.86
Binary Dirichlet	20.61	26.64	0.7389	53.12
CHAIN	**8.53**	**20.18**	**0.5904**	**69.47**

Table 15: LLM forecasting-call control and multi-sample Direct-MO comparison with DeepSeek-V3 across eight datasets. ECE, Brier, and Accuracy are in %.

Method	LLM calls/case	Tokens/case	ECE↓	Brier↓	NLL↓	Acc↑
Direct-MO	1.00	550	15.21	23.47	0.6627	60.25
Five-sample Direct-MO	5.00	2,751	14.91	23.07	0.6514	60.84
CHAIN	5.00	13,371	**8.14**	**19.89**	**0.5835**	**70.02**

each dataset and the same metric definitions as in the main text; LLM calls and token consumption are computed from the corresponding inference runs.

Increasing Direct-MO from one to five independent forecasts reduces ECE from 15.21 to 14.91, Brier score from 23.47 to 23.07, and NLL from 0.6627 to 0.6514, while increasing Accuracy from 60.25 to 60.84. Independent repeated forecasting and probability averaging therefore produce only small improvements across all four metrics, and their overall performance remains below that of **CHAIN**.

Relative to five-sample Direct-MO, **CHAIN** reduces ECE and Brier score by 6.77 and 3.18 percentage points, respectively, reduces NLL by 0.0678, and improves Accuracy by 9.18 percentage points. Both methods use approximately 5.00 LLM calls per evaluation instance, while **CHAIN** performs better on all four predictive metrics. These results show that increasing the number of independent forecasts and averaging their probabilities does not attain the gains of **CHAIN**; with essentially the same number of LLM calls, structured causal-temporal inference more effectively improves probability calibration and predictive accuracy.

M Robustness Across Calibration Error Estimators

To test whether CHAIN's calibration advantage depends on a particular error estimator, we evaluate the same predictive outputs with four estimators that use different binning and finite-sample bias treatments: Adaptive ECE-10, debiased RMSCE, Plugin RMSCE, and Jackknife bias-corrected ECE. All metrics are computed from the same verified set of test-set predictions. The evaluation comprises 24 evaluated dataset–backbone combinations spanning 8 datasets and 3 backbone language models, and Table 16 reports the macro-average over these combinations.

Across all four calibration error estimators, **CHAIN** obtains lower error than LLM Direct. Specifically, its Adaptive ECE is 7.402 percentage points lower, its debiased RMSCE is 12.617 percentage points lower, its Plugin RMSCE is 9.165 percentage points lower, and its Jackknife bias-corrected ECE is 11.295 percentage points lower.

The four estimators characterize calibration from complementary perspectives. Adaptive ECE-10 uses equal-frequency bins to reduce sensitivity to uneven predicted-probability density; debiased RMSCE corrects the finite-sample bias in squared calibration error; Plugin RMSCE directly estimates root mean squared calibration error from empirical bin statistics; and Jackknife bias-corrected ECE corrects finite-sample bias through leave-one-out resampling. The consistently lower errors of **CHAIN** show that its calibration improvement is not an artifact of a particular binning scheme or bias-correction procedure.

Table 16: Macro-averaged calibration errors under four estimators across 24 evaluated dataset–backbone combinations spanning 8 datasets and 3 backbone language models.

Method	Adaptive ECE-10↓	Debiased RMSCE↓	Plugin RMSCE↓	Jackknife ECE↓
LLM Direct	19.144	16.654	20.639	15.606
CHAIN	**11.742**	**4.037**	**11.474**	**4.312**

Table 17: Extended calibration diagnostics and predictive performance, macro-averaged over 24 dataset–backbone combinations. All metrics except NLL are reported as their original values multiplied by 100.

Method	MCE-10↓	Reliability-10↓	Resolution-10↑	Brier↓	NLL↓	Acc.↑
LLM Direct	39.121	4.530	2.805	25.314	0.721	61.491
CHAIN	**22.701**	**1.428**	**4.716**	**20.178**	**0.590**	**69.466**

N EXTENDED CALIBRATION DIAGNOSTICS AND PROBABILITY RESOLUTION

We further assess probabilistic prediction quality through complementary measures of maximum bin-wise calibration error, reliability, probability resolution, strictly proper scoring rules, and classification accuracy. All metrics are computed from the same verified test-set predictions used in the preceding section. The analysis covers 24 dataset–backbone combinations formed by 3 backbone language models and 8 datasets, with Table 17 reporting the macro-average over these combinations. MCE-10 is the largest absolute gap between mean predicted event probability and empirical event frequency among ten probability bins. Reliability-10 measures the weighted squared calibration gap across bins, whereas Resolution-10 measures between-bin variation in empirical outcome frequencies. We also report Brier score, negative log-likelihood (NLL), and Accuracy to jointly evaluate probabilistic predictions and classification decisions.

As shown in Table 17, **CHAIN** outperforms LLM Direct on all six complementary metrics. It reduces MCE-10 from 39.121 to 22.701 and Reliability-10 from 4.530 to 1.428, corresponding to reductions of 16.420 and 3.102 on the reported $\times 100$ scale, respectively. These results show that **CHAIN** reduces both the maximum bin-wise calibration error and the aggregate reliability error. Resolution-10 increases from 2.805 to 4.716, indicating that its predicted probabilities more effectively distinguish samples with different empirical outcome frequencies.

For overall probabilistic prediction quality, **CHAIN** reduces Brier score from 25.314 to 20.178 and NLL from 0.721 to 0.590, while increasing Accuracy from 61.491 to 69.466. The concurrent improvements in calibration error, probability resolution, strictly proper scoring rules, and classification accuracy demonstrate that **CHAIN** produces probabilities that better align with empirical outcomes while providing stronger discrimination and predictive performance.

O CAUSAL-COVERAGE RESPONSE OF THE ADAPTIVE FUSION WEIGHT

To characterize how **CHAIN** adjusts the contributions of its prediction sources according to question-specific causal evidence, we analyze the causal coverage $\Omega(q)$ and final fusion weight $\alpha(q)$ for DeepSeek-V3 on the evaluation sets of all eight datasets. Causal coverage summarizes the number of retained causal chains, their mean confidence, evidence reliability, and directional balance. The fusion weight $\alpha(q)$ determines the contribution of the causal estimate $P_{\mathrm{causal}}(q)$ to the final forecast. The recorded weights use mapping scale $\alpha_b = 0.5$, fusion upper bound $\alpha_0 = 0.6$, coverage activation threshold $\Omega_0 = 0.3$, and sigmoid-transition sharpness $\zeta = 3$; when the retained-chain count is $k = 0$, the zero-chain gate sets $\alpha(q) = 0$. Table 18 stratifies the verified prediction records by causal coverage and reports the share of records and the mean and median fusion weights within each stratum.

As shown in Table 18, the fusion weight increases with causal coverage. In the prediction records, the $\Omega(q) = 0$ stratum corresponds to $k = 0$; the zero-chain gate therefore sets the recorded fusion weight to zero, and the final forecast is given by $P_{\mathrm{llm}}(q)$. For nonzero but low causal coverage, the mean fusion weights are 0.309 and 0.399 in the first two positive-coverage strata. The mean

Table 18: Adaptive fusion weights stratified by causal coverage on DeepSeek-V3 across the evaluation sets of all eight datasets.

Causal coverage $\Omega(q)$	**Record share (%)**	**Mean** $\alpha(q)$	**Median** $\alpha(q)$
$\Omega(q) = 0$	49.22	0.000	0.000
$0 < \Omega(q) < 0.10$	22.46	0.309	0.305
$0.10 \leq \Omega(q) < 0.25$	7.42	0.399	0.404
$0.25 \leq \Omega(q) < 0.50$	3.22	0.541	0.536
$0.50 \leq \Omega(q) \leq 1.00$	17.68	0.600	0.600

Table 19: Definition of the node-level causal-distance evidence audit.

Component	**Evidence-occurrence weighted**	**Case-balanced**
Analysis unit	Each occurrence of a non-target evidence node in a retained chain	Each case–distance stratum
Distance strata	$d(n, q) \in \{1, \ldots, d_{\max}\}$	$d(n, q) \in \{1, \ldots, d_{\max}\}$
Aggregation	Mean $V_{\mathrm{CTVF}}(n, q)$ across node occurrences at each distance	Mean within each case and distance, followed by an average across cases
Target handling	Exclude $n \in \mathcal{T}(q)$	Exclude $n \in \mathcal{T}(q)$

weight increases to 0.541 for $0.25 \leq \Omega(q) < 0.50$. When causal coverage reaches 0.50 or above, all corresponding prediction records attain the evaluation-time upper bound of 0.600. Across all prediction records, the Spearman rank correlation between $\Omega(q)$ and $\alpha(q)$ is 0.998, summarizing their monotonic association in the recorded forecasts.

These results verify the expected response of $\alpha(q)$ to question-specific causal coverage: the causal estimate does not contribute to the final forecast when no usable causal evidence is available, its contribution increases as evidence coverage grows, and it reaches the prescribed upper bound in the high-coverage regime. Thus, **CHAIN** does not apply a single fixed fusion weight, but adjusts the contribution of the causal estimate to the final probability within a bounded range according to question-level causal coverage.

P CAUSAL-DISTANCE-STRATIFIED EVIDENCE-NODE AUDIT PROTOCOL

To characterize how **CHAIN** assigns validity across causal distances, we define a node-level audit over the non-target evidence nodes appearing in retained causal chains. Every retained chain terminates at a target entity, so its terminal anchor has distance zero by construction and is excluded from the audit. For each remaining node $n \notin \mathcal{T}(q)$, we record its causal distance $d(n, q)$ and its assigned validity $V_{\mathrm{CTVF}}(n, q)$, and stratify the observations by $d(n, q) \in \{1, \ldots, d_{\max}\}$. Table 19 formalizes the two complementary aggregation schemes.

The evidence-occurrence-weighted estimate measures the validity assigned to all recorded non-target node occurrences at a given distance. The case-balanced estimate first averages within each case and distance stratum and then averages across cases, preventing cases with more retained chains from dominating the comparison. This formulation is consistent with the target-terminated chain pool and isolates the distance–validity relationship without conflating it with the mechanically zero distance of the terminal target anchor.

Q CONTROLLED PERTURBATIONS OF CAUSAL-EDGE DIRECTION AND STRENGTH ORDERING

To isolate the effects of causal-edge direction and relative strength ordering, we conduct a controlled perturbation study with DeepSeek-V3 on all eight datasets. We hold fixed the forecasting questions, target entities, node relevance scores, language-model probabilities, adaptive fusion rule, and search settings, modifying only causal-edge direction or the relative ordering of edge strengths. Causal distances, retrieval chains, and final predictions are recomputed under every setting. The three perturbations randomly permute strengths among causal edges, apply the complementary transformation $w \leftarrow 1 - w$, or reverse all causal-edge directions.

Table 20: Macro-average performance changes across eight datasets under causal-edge direction and strength-order perturbations with DeepSeek-V3, relative to the unperturbed configuration. ECE, Brier, and Accuracy changes are in percentage points.

Perturbation	ΔECE↓	ΔBrier↓	ΔAccuracy↑	ΔNLL↓
Unperturbed CHAIN	0.000	0.000	0.000	0.00000
Shuffle edge strengths	-0.034	$+0.024$	-0.098	$+0.00033$
Complement strengths ($w \leftarrow 1 - w$)	$+0.829$	$+0.015$	-0.195	-0.00003
Reverse causal-edge directions	$+1.689$	$+1.996$	-5.371	$+0.06869$

Table 20 reports macro-average metric changes relative to the unperturbed configuration, defined as $\Delta m = m_{\text{perturbed}} - m_{\text{original}}$. Positive changes in ECE, Brier score, and NLL indicate degradation, whereas a negative change in Accuracy indicates degradation. Changes in ECE, Brier score, and Accuracy are measured in percentage points.

As shown in Table 20, reversing causal-edge directions increases ECE and Brier score by 1.689 and 1.996 percentage points, decreases Accuracy by 5.371 percentage points, and increases NLL by 0.06869. All four metrics change in the direction of degradation, and their absolute changes are the largest among the three perturbations. Direction reversal preserves the nodes, undirected adjacency, and edge strengths while changing only edge orientation. The concurrent deterioration in calibration error, probabilistic loss, classification accuracy, and negative log likelihood therefore shows that the evidence-propagation order and search reachability induced by causal-edge direction provide important structural information for organizing **CHAIN**'s inference paths.

Randomly shuffling edge strengths preserves the original multiset and overall distribution of strengths but changes their association with individual causal edges. This perturbation changes ECE by -0.034, Brier score by $+0.024$, Accuracy by -0.098, and NLL by $+0.00033$. The small magnitudes show that local reassignment of edge strengths does not substantially alter overall predictive performance.

The complementary transformation further reverses the relative ordering of the original strengths. It increases ECE by 0.829 percentage points and decreases Accuracy by 0.195 percentage points, while changing Brier score and NLL by only $+0.015$ and -0.00003, respectively. Compared with random shuffling, this systematic reversal has a larger effect on calibration and classification decisions, but every change remains smaller than that induced by reversing causal-edge directions. Thus, strength ordering modulates the relative contribution of inference paths, while **CHAIN** remains comparatively stable across the two strength-order perturbations.

Overall, direction and strength-order perturbations produce distinct performance responses. Strength shuffling and complementation yield limited changes with mixed signs, whereas direction reversal causes a consistent and largest degradation across all four metrics. Under this controlled evaluation, causal-edge direction has a more consistent effect on predictions than the particular ordering of edge strengths, providing empirical support for direction-sensitive causal retrieval while demonstrating robustness to changes in strength ordering.

R DOMAIN-SPECIFIC RETRIEVAL-CORPUS CONTROL

R.1 CORPUS CONSTRUCTION AND BOUNDARY AUDIT

We conduct a domain-specific retrieval-corpus control to test whether **CHAIN**'s relative performance depends on a mismatch between the retrieval corpus and the target data source.

For Clinical Trial Outcomes, FOReCAst, Golf-Forecasting, and KalshiBench, we construct target-domain-relevant corpora with dataset-level temporal cutoffs, retaining only documents available on or before the corresponding cutoff. Documents are collected from prespecified domain categories rather than through question-directed retrieval; evaluation-question records and fields containing answers, labels, or resolution information are excluded. Within each dataset, NaiveRAG, GraphRAG, HyperGraphRAG, and **CHAIN** share the same audited source-document set, cutoff, and exclusion rule while constructing their native retrieval structures. Table 21 summarizes the corpus boundaries.

Table 21: Corpus-boundary audit for the domain-specific retrieval-corpus control.

Dataset	Common cutoff	Documents	Chunks	Test-record exclusion	Shared boundary
Clinical Trial Outcomes	2024-12-20	11	24	Yes	Identical
FOReCAst	2023-01-01	9	23	Yes	Identical
Golf-Forecasting	2025-10-12	9	27	Yes	Identical
KalshiBench	2025-08-19	12	39	Yes	Identical

Table 22: Macro-averaged performance under the domain-specific retrieval-corpus control over four OOD datasets with DeepSeek-V3.

Method	ECE↓	Brier↓	NLL↓	Acc↑
NaiveRAG	12.71	23.45	0.663	60.94
GraphRAG	17.57	24.15	0.705	59.57
HyperGraphRAG	12.19	22.95	0.650	62.70
CHAIN	**8.22**	**22.57**	**0.640**	**63.09**

All four corpus audits return `status=passed` with empty error lists. Retained documents have parseable dates, satisfy the corresponding cutoff, and exclude evaluation-question records. Within each dataset, all methods are bound to the same source-document set, holding corpus scope, temporal boundaries, and exclusion conditions fixed across methods.

R.2 Performance under Domain-Specific Retrieval Corpora

Using DeepSeek-V3, we evaluate NaiveRAG, GraphRAG, HyperGraphRAG, and **CHAIN** on the same test questions and metric definitions. Table 22 reports equal-weight macro averages over the four OOD datasets.

Table 22 shows that **CHAIN** achieves the best macro average on all four metrics. Relative to the metric-specific strongest retrieval baseline, it lowers ECE and Brier by 3.964 and 0.381 percentage points, respectively, lowers NLL by 0.010, and increases Accuracy by 0.391 percentage points. These macro-average results show that its overall calibration and predictive advantages persist when all methods share the same corpus scope, temporal boundaries, and exclusion conditions.

S Case Study

As shown in Figure 11, for the question of whether IBM would officially launch an AI-sovereignty product or service marketed under the name "Sovereign Core" by February 28, 2026, all eleven baselines shown in the figure (LLM Direct, five RAG methods, and five post-hoc calibrators) produce a *no* prediction at confidences ranging from 53.75% to 99.00%, whereas the ground truth is *yes*. Among the methods shown in Figure 11, only **CHAIN** produces the correct *yes* prediction. Its Brier score is 2.56%, compared with 28.89%–98.01% for the displayed baselines, and is therefore approximately $1/11.3$–$1/38.3$ of the corresponding baseline scores. Its absolute probability error $100|p-y|$ is 16.00, compared with 53.75–99.00 for the displayed baselines, or approximately $1/3.4$–$1/6.2$ of their errors. Table 23 further reports $P(\text{yes})$ at each stage inside **CHAIN**, showing how directionally consistent causal evidence strengthens the probability support for an existing positive prediction.

The displayed baselines exhibit a consistent prior-dominated failure. LLM Direct rejects the event because there is no direct evidence that IBM had announced a product named "Sovereign Core." Although the retrieval baselines recover information related to IBM's AI and cloud services, the retrieved context does not explicitly mention "Sovereign Core," and the absence of a direct reference is still interpreted as evidence against the event. The five post-hoc calibrators transform the probability produced by LLM Direct, but none corrects the prediction direction in this case.

CHAIN obtains the correct, lower-error forecast through two coupled structural stages. Anchored at the IBM entity in the causal-temporal hypergraph, direction-aware evidence aggregation retains 39 valid causal chains, of which 34 support and 5 oppose the event, with causal-evidence coverage of 1.00. Direction-aware Noisy-OR aggregates these chains into $P_{\text{causal}} = 0.99999$, providing stronger

Question	Will IBM officially launch an AI sovereignty product or service marketed under the name 'Sovereign Core' by February 28, 2026?														
Ground Truth	yes														
Method	LLM Direct					NaiveRAG					GraphRAG				
Reasoning Summary	IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.					The provided context does not mention IBM or any 'Sovereign Core' product/service, and there is no relevant information to suggest such a launch. Given the lack of evidence, it is unlikely.					There is no specific context provided about IBM's plans for an AI sovereignty product named 'Sovereign Core'. However, IBM has a history of launching AI and cloud-related services, but without explicit evidence or announcements, the likelihood remains speculative.				
Metric	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC
Scores	no	65.00	65.00	42.25	0.00	no	65.00	65.00	42.25	0.00	no	65.00	65.00	42.25	0.00
Method	LightRAG					HippoRAG					HyperGraphRAG				
Reasoning Summary	The context does not mention any IBM product or service named 'Sovereign Core' or any plans for such a launch by February 28, 2026. While IBM is active in AI and related initiatives, there is no evidence to suggest this specific product or service will be launched.					The provided context does not mention any IBM initiatives related to 'Sovereign Core' or AI sovereignty products. IBM's recent focus, as indicated in the context, has been on AI development, acquisitions, and platform enhancements like watsonx Orchestrate and watsonx.data intelligence. Without specific evidence or announcements pointing towards a 'Sovereign Core' product, it is unlikely to be launched by the specified date.					The context provided does not mention any product or service named 'Sovereign Core' or any plans related to AI sovereignty by IBM. While IBM is actively involved in AI development, there is no evidence to suggest the launch of such a specific product by February 28, 2026.				
Metric	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC
Scores	no	70.00	70.00	49.00	0.00	no	65.00	65.00	42.25	0.00	no	65.00	65.00	42.25	0.00
Method	Temperature Scaling					Platt Scaling					Isotonic Regression				
Reasoning Summary	IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.					IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.					IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.				
Metric	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC
Scores	no	60.51	60.51	36.61	0.00	no	79.25	79.25	62.81	0.00	no	82.35	82.35	67.82	0.00
Method	Histogram Binning					Conformal-score Adj.					**CHAIN (Ours)**				
Reasoning Summary	IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.					IBM has not announced any plans for a product named 'Sovereign Core', and the timeline is speculative.					The retrieved causal evidence consistently supports IBM's expansion into sovereign AI offerings, making an official launch under the name 'Sovereign Core' by the specified date likely.				
Metric	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC	Answer	Confidence	100|p−y|	Brier	ACC
Scores	no	99.00	99.00	98.01	0.00	no	53.75	53.75	28.89	0.00	**yes**	**84.00**	**16.00**	**2.56**	**100.00**

Figure 11: Single-question forecast comparison for a resolved forecasting question, comparing LLM Direct, five RAG methods (NaiveRAG, GraphRAG, LightRAG, HippoRAG, HyperGraphRAG), and five post-hoc calibrators (Temperature Scaling, Platt Scaling, Isotonic Regression, Histogram Binning, Conformal-Score Probability Adjustment) based on DeepSeek-V3, with **CHAIN** (Ours). The displayed $100|p - y|$ is the absolute probability error for this instance, not dataset-level ECE.

Table 23: External reference prediction and internal probability components of **CHAIN** in the case study (ground truth: *yes*, $P^\star = 1$). The first row provides the standalone LLM Direct reference; the remaining rows report CHAIN's LLM-side estimate, causal estimate, and final adaptive-fusion output.

Prediction source	$\boldsymbol{P}$**(yes)**	$\boldsymbol{\lvert P - P^\star \rvert}$
LLM Direct (external reference)	0.3500	0.6500
P_{llm} (LLM inside CHAIN)	0.6000	0.4000
P_{causal} (Noisy-OR over 39 chains)	0.99999	0.00001
$\hat{p}$ (Adaptive Fusion, $\alpha = 0.60$)	**0.8400**	**0.1600**

causal probability support for the existing positive LLM-side forecast. Causal-Coverage-Balanced Adaptive Fusion then combines $P_{\mathrm{llm}} = 0.6000$ and $P_{\mathrm{causal}} = 0.99999$ with $\alpha = 0.60$, producing $\hat{p} = 0.8400$, which is the 84.00% confidence reported in Figure 11.

As Table 23 shows, standalone LLM Direct assigns only 0.3500 probability to the true outcome, whereas CHAIN's internal LLM-side estimate is $P_{\mathrm{llm}} = 0.6000$. Within CHAIN, direction-aware causal aggregation produces $P_{\mathrm{causal}} = 0.99999$ from the 39 valid causal chains, and adaptive fusion combines it with P_{llm} to yield $\hat{p} = 0.8400$. The final probability retains the uncertainty represented by the LLM-side estimate while using directionally consistent causal evidence to increase confidence in the positive prediction, resulting in lower Brier score and absolute probability error than all eleven baselines shown in Figure 11.

Table 24: Configuration-level paired comparisons of **CHAIN** with the metric-specific strongest baseline across 24 (dataset, backbone) configurations spanning three backbone LLMs and eight forecasting datasets. All tests are one-sided with the alternative hypothesis "**CHAIN** better" (ΔECE, ΔBrier < 0; ΔAcc > 0); $n = 24$. Median differences are in percentage points.

Metric	Median Δ	Wilcoxon p	Paired-t p	Wins	Binomial p
ΔECE$\downarrow$	-0.581	$\mathbf{5.96\times10^{-8}}$	$\mathbf{5.46\times10^{-4}}$	$24/24$	$\mathbf{5.96\times10^{-8}}$
ΔBrier$\downarrow$	-1.765	$\mathbf{5.96\times10^{-8}}$	$\mathbf{7.37\times10^{-7}}$	$24/24$	$\mathbf{5.96\times10^{-8}}$
ΔAcc$\uparrow$	$+2.734$	$\mathbf{5.96\times10^{-8}}$	$\mathbf{4.34\times10^{-7}}$	$24/24$	$\mathbf{5.96\times10^{-8}}$

Table 25: Question-level paired comparisons between **CHAIN** and five retrieval-augmented baselines across three backbone LLMs and eight datasets. Δ denotes **CHAIN** minus the baseline in percentage points; the CI column counts paired comparisons whose 95% bootstrap interval lies entirely in the favorable direction.

Metric	Favorable comparisons	Median Δ (pp)	Favorable 95% CI	Raw $p < .05$	Holm $p < .05$
ECE $\downarrow$	**120/120**	-9.2591	49/120	89/120	44/120
Brier $\downarrow$	**120/120**	-5.7738	75/120	84/120	27/120
Accuracy $\uparrow$	**120/120**	$+12.1094$	58/120	67/120	20/120

T CONFIGURATION-LEVEL COMPARATIVE ANALYSIS

To summarize **CHAIN**'s empirical improvements over the metric-specific strongest baseline, we conduct a paired analysis on the full matrix of 24 (dataset, backbone) configurations spanning three backbone large language models and eight forecasting datasets, treating each configuration as one paired observation. For each configuration and each evaluation metric, the strongest baseline is identified separately by the best per-metric value among the thirteen baselines under comparison: lowest ECE for ECE, lowest Brier score for Brier, and highest Accuracy for Accuracy.

We report three complementary one-sided tests on the resulting paired differences: the non-parametric Wilcoxon signed-rank test, the parametric paired t-test, and the binomial test on the number of configurations in which **CHAIN** outperforms the metric-specific strongest baseline. The one-sided alternatives favor lower ECE and Brier score or higher Accuracy for **CHAIN**.

As shown in Table 24, the selected metric-specific strongest baseline is outperformed by **CHAIN** in all 24 configurations for each metric. For ECE, Brier score, and Accuracy, **CHAIN** outperforms the metric-specific strongest baseline in all 24 configurations. The median paired differences are -0.581 percentage points for ECE, -1.765 percentage points for Brier score, and $+2.734$ percentage points for Accuracy. The agreement among the Wilcoxon signed-rank test, paired t-test, and configuration-level win-count test is consistent with these favorable paired differences. Because the comparison baseline is selected from the same evaluation results, the reported tests are descriptive rather than confirmatory significance tests.

U QUESTION-LEVEL PAIRED TESTS

To assess differences at the question level, we pair **CHAIN** with NaiveRAG, GraphRAG, LightRAG, HippoRAG, and HyperGraphRAG on identical test questions and labels for GPT-4o-mini, Gemini-2.0-Flash, and DeepSeek-V3 across all eight datasets, yielding 120 paired comparisons. Each comparison corresponds to one (backbone, dataset, baseline) combination.

For ECE, we use a paired prediction-swap randomization test with 10,000 randomizations; for Brier score, we use a paired sign-flip permutation test with 10,000 randomizations. For Accuracy, we use the exact McNemar test. We additionally construct 95% paired bootstrap intervals for all three metrics using 10,000 resamples. All tests use prespecified one-sided alternatives that favor lower ECE and Brier score or higher Accuracy for **CHAIN**. Holm correction is applied separately to the 120 tests for each metric to control the family-wise error rate. Table 25 reports directional consistency, median effect size, bootstrap support, and significance counts before and after multiple-comparison correction.

As shown in Table 25, **CHAIN** has the favorable direction in all 120 paired comparisons for each metric: it yields lower ECE and Brier score and higher Accuracy for every (backbone, dataset, baseline) combination. This directional consistency spans all three backbone LLMs, all eight datasets, and all five retrieval-augmented baselines. The median paired differences also favor **CHAIN**: ECE and Brier score decrease by 9.2591 and 5.7738 percentage points, respectively, while Accuracy increases by 12.1094 percentage points.

Before multiple-comparison correction, 89 ECE, 84 Brier, and 67 Accuracy comparisons attain $p < .05$. After Holm correction, 44, 27, and 20 comparisons, respectively, remain significant, with minimum adjusted p-values of 0.0120, 0.0120, and 4.08×10^{-8}. The 95% bootstrap interval lies entirely in the favorable direction in 49 ECE, 75 Brier, and 58 Accuracy comparisons. The unanimous direction of the paired effects, their median magnitudes, the bootstrap intervals, and the Holm-adjusted tests provide mutually reinforcing evidence that **CHAIN** consistently improves calibration, probabilistic prediction quality, and classification accuracy across backbone models, datasets, and retrieval-augmented baselines.

V APPLICATION ANALYSIS

Our evaluation covers financial events and event-contract markets (Polymarket and KalshiBench), technology and industry (AI-Futures), clinical research and development (Clinical Trial Outcomes), sports forecasting (Golf-Forecasting), weather and climate events, and general open-domain event forecasting (Metaculus, Future-as-Label, and FOReCAst). Weather- and climate-related questions occur in Metaculus, Future-as-Label, FOReCAst, and KalshiBench. In these domains, the practical value of probabilistic forecasting depends not only on directional accuracy, but also on whether predicted probabilities retain a consistent statistical meaning across questions and information environments. Systematic miscalibration causes risk thresholds, event rankings, and resource-allocation decisions to be based on biased confidence. **CHAIN** maps heterogeneous, dynamic, and potentially dependent evidence into calibrated probabilities through causal-temporal evidence weighting, direction-aware aggregation, and causal-coverage-adaptive fusion. Its consistent advantages across domains and under distribution shift provide methodological and empirical support for the applications discussed below.

Financial Events and Event-Contract Markets. Event-contract and macro-risk analysis require probabilities that are directly comparable across events and can support exposure ranking, threshold management, and scenario monitoring. These information environments often contain reports that are source-correlated or semantically duplicated; treating them as independent evidence can systematically inflate confidence. **CHAIN**'s direction-aware deduplication and evidence aggregation control repeated accumulation while retaining the structure of evidence both for and against an event. Causal-coverage-adaptive fusion further regulates the contribution of the structured estimate according to effective causal coverage. The resulting probabilities can support event-level risk assessment, contract screening, and macro-event monitoring.

Technology Development and Public-Event Analysis. Technology milestones, industrial evolution, and public events are commonly driven by multiple interacting factors whose evidence differs in temporal validity and causal proximity. **CHAIN** assigns evidence weights according to event-time differences and causal distance, increasing the contribution of recent evidence that is more directly connected to the forecast target while reducing the influence of stale background information and distant associations. This structure can support technology-roadmap monitoring, industrial scenario comparison, critical-event alerts, and probability updates at decision points.

Clinical Trials and Drug Development. Clinical-development decisions require the integration of trial design, drug mechanisms, disease context, interim outcomes, and external research evidence to continuously assess project-level success probabilities. Such evidence often exhibits explicit causal dependencies and temporal ordering that are not preserved by simply concatenating retrieved passages. **CHAIN** organizes relevant evidence into traceable causal chains and derives the final probability from chain reliability, directional balance, and causal coverage. These event-level probabilities can support trial-risk assessment, development-pipeline screening, candidate prioritization, and resource allocation on a common probabilistic scale.

Weather and Climate Event Forecasting. Weather and climate event forecasting requires the integration of historical climate conditions, recent observations, seasonal variation, and continuously updated event information into probabilities with a clear statistical meaning. Such evidence is strongly time-dependent and may be repeated across multiple sources. **CHAIN**'s temporal-validity weighting reduces the influence of stale information, direction-aware deduplication limits repeated accumulation of duplicated reports, and causal-distance weighting together with causal-coverage-adaptive fusion regulates the final probability according to the relevance of the evidence to the forecast target. These mechanisms can support probabilistic forecasts, risk monitoring, and warning decisions for weather and climate events such as rainfall thresholds, storm and hurricane occurrence, extreme temperatures, snowfall, and drought.

Sports Forecasting. Sports events have short resolution horizons, frequent information updates, and a strong dependence on recent conditions. A forecasting system must incorporate new evidence promptly while preventing historical information or repeated reporting from exerting disproportionate influence. The temporal decay in **CHAIN** reduces the weight of less timely evidence, causal-distance weighting prioritizes information more closely connected to the target outcome, and direction-aware deduplication limits confidence shifts caused by repeated descriptions of the same event. The resulting estimates can support event-state monitoring, outcome-probability updating, and critical-event alerts.

Open-Domain Event Forecasting. Open-domain forecasting spans heterogeneous topics, evidence sources, and resolution horizons, requiring probability quality to remain stable as the domain and information distribution change. **CHAIN** does not depend on a fixed set of domain-specific features; instead, it constructs each forecast from question-conditioned causal structure, temporal validity, and evidence coverage. Its cross-domain and distribution-shift results support its use as a general probabilistic inference layer for open-domain forecasting platforms, risk-monitoring systems, and analyst workflows, providing outputs with a consistent statistical meaning for question ranking, scenario tracking, threshold-based decisions, and human review.

W LIMITATIONS

While **CHAIN** shows consistent calibration advantages in our experiments, five limitations remain. First, the causal-temporal hypergraph is constructed by an LLM extractor, so errors in entities, relation directions, or strengths can propagate to downstream aggregation and fusion; deduplication and directional-balance correction mitigate but do not eliminate this uncertainty. Second, bounded reachability and finite chain-enumeration budgets trade runtime efficiency against long-range coverage and may omit relevant evidence. Third, the current output is an endpoint-event probability rather than an explicit model of intermediate states, transition probabilities, event timing, or alternative trajectories. Fourth, per-query inference requires multiple LLM calls and a non-trivial prompt budget, which may constrain large-scale or low-latency deployment. Fifth, temporal admissibility depends on corpus boundaries and metadata, and empirical validation remains concentrated on English event-forecasting benchmarks and selected knowledge-intensive domains. These limitations concern robustness, computational scaling, task scope, and external validity rather than the core calibration mechanism under the evaluated forecasting protocol.

X FUTURE WORK

Future work will address these limitations in five directions. First, multi-sample extraction, cross-chunk consistency constraints, and outcome-supervised relation estimation can improve the hypergraph and propagate extraction uncertainty into aggregation and fusion. Second, reachability and chain-enumeration budgets can become query-adaptive, allocating deeper search when evidence is distant and stopping when sufficient directional evidence has accumulated. Third, **CHAIN** can extend endpoint prediction to process-level event trajectories covering intermediate states, transitions, timing, and alternative paths. Fourth, prompt compression, auxiliary-model distillation, subgraph caching, and reuse of stable relation judgments can reduce inference cost. Fifth, joint temporal-boundary validation and evaluation on multilingual, scientific, legal, and low-resource forecasting can broaden evidence for calibration, temporal validity, and interpretable decision support. These extensions will test whether the same mechanisms remain reliable under broader evidence, languages, temporal boundaries, and deployment constraints and broader operating conditions.