Outcome-Free Audits and Repairs for LLM Forecasters

Anonymous Authors[1]

Abstract

LLM forecasters now post Brier scores rivaling crowd aggregates on ForecastBench and Prophet Arena, yet their probabilities can be exploited by a Dutch book that requires no resolved outcomes; the questions that matter most resolve years after deployment, where proper-scoring-rule loss is silent and unavailable. We introduce a *deployment-time audit stack*: given a forecast vector, it detects violations of probability coherence and reports four classes of Dutch-book exposure (complementary pairs, monotonicity chains, Fréchet conjunction bounds, logical entailment) without waiting for resolution. For complementary pairs the symmetric coherent projection has an exact Brier-improvement identity (Theorem 1), so the audit is therefore repairable. We release **CoherenceBench v0** (262 questions, 68 pairs, 19 chains, 39 conjunction triples, 21 entailments) with the audit harness.[1] Coherence audits complement rather than replace proper-scoring-rule evaluation: a forecaster can be perfectly coherent and uninformative ($f \equiv 0.5$) or accurate and incoherent. Across 15 contemporary forecasters (9 open-weight, 6 closed),[2] no observed forecaster is violation-free on all four axes (zeros on small-n axes upper-bounded, not conclusive), and the model with the best aggregate coherence profile (R1-Distill-Qwen-32B-AWQ) has $\sim 15\times$ the mean Brier of GPT-5 (0.27 vs 0.018). Symmetric coherent projection recovers 21.09% of pair-Brier loss without additional model calls; per-pair improvement equals $g^2/4$ for $g = f(A) + f(B) - 1$ (identity verified to numerical precision on 991 labeled complementary observations).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Anonymized release: https://anonymous.4open.science/r/coherencebench-v0-C384/

[2]Closed-model identifiers refer to API endpoints; exact IDs, request parameters, system prompt, JSON schema, retry policy, and raw response logs ship in Appendix F.

1. Introduction

Prediction-market platforms (Kalshi, Polymarket, Metaculus) and research benchmarks (Karger et al., 2024; Halawi et al., 2024) routinely deploy LLM forecasters on questions that resolve months or years out: AGI timelines, election cycles, geopolitical onset conditions. Resolved-outcome evaluation is structurally unavailable on these questions at the moment a deployment decision is made, and proper-scoring-rule loss (when it eventually arrives) is silent on whether the model's probabilities respect basic axioms of probability theory across related events. CoherenceBench is the audit a deployment engineer runs before approving an LLM forecaster, on the same forecast vector that question solicitation already produces.

Every probabilistic forecast is also a self-disclosure: if a model forecasts A and the disjoint A^c, the two outputs jointly satisfy or violate $P(A)+P(A^c)=1$ regardless of how either question resolves. Coherent forecasting requires the constraint to hold; the two outputs are not independent samples but joint outputs of a single generative model on the same date and the same prompt template. We exploit this self-disclosure to construct four outcome-free Dutch-book audits over structured event groups (complementary pairs, monotonicity chains, Fréchet conjunction bounds, and logical entailment) that detect deterministic violations of probability axioms in any forecast set already collected. No additional inference is required, and because each audit indicator is a function of the forecast vector alone the rates are unaffected by whether the underlying event has leaked into pre-training: a model that has seen the Macron resolution still pays the same arbitrage if its $P(\text{Macron})+P(\text{Le Pen})$ misses one. By *outcome-free* we mean no audit decision requires outcomes; resolved Brier is reported alongside as a downstream cross-check.

Across 15 contemporary forecasters on CoherenceBench v0 (262 events; nine open-weight, six closed) the four-axis violation rates are $68.0\%, 11.7\%, 35.3\%$, and 15.4% (Table 2); no model is clean on all four. The $n=1005$ pair observations exceed the anytime-valid SPRT bound for rejecting $p \le 0.5$ by an order of magnitude (Appendix E). R1-Distill-Qwen-32B-AWQ (an open-weight reasoning-distilled model) has the best aggregate coherence profile, yet its mean Brier is roughly $15\times$ GPT-5's (0.27 vs 0.018): across the 15 fore-

casters, capability (Brier) and coherence (Dutch-book violation rates) rank orthogonally. Symmetric coherent projection on each pair recovers 21.09% of pair-Brier loss without additional model calls, with the per-pair improvement equal to exactly $g^2/4$ (Theorem 1; empirical residual $< 10^{-5}$).

Contributions. The integrated stack: **(i)** a four-axis outcome-free audit framework that scores any forecast vector the day it is collected (§3); **(ii)** Theorem 1, an identity exhibiting a constructive lower bound on the proper-scoring-rule loss recoverable from any coherent pair, with the symmetric projection as the witness operator. The algebra is straightforward; the bound is what makes the audit actionable at deployment time; **(iii)** a deployable repair layer any forecaster can wrap with no retraining (§4); **(iv)** the released benchmark and audit harness that operationalize (i)–(iii) on 15 contemporary models. Coherence is complementary to, not a substitute for, calibration.

2. Related Work

Table 1 positions this work against the closest prior art on *probabilistic* coherence: Zhu et al. (2024) test pair-complementarity on crowdsourced binaries with no proper-scoring link or repair; Liang et al. (2025) project onto the coherent simplex but supply no scoring-rule identity or benchmark; Irugalbandara et al. (2025) filter forecasts by *argumentative* coherence (rhetorical consistency of supporting arguments) instead. Upstream foundations on LLM calibration (Kadavath et al., 2022; Tian et al., 2023) and output-consistency (Elazar et al., 2021; Mündler et al., 2024) and on internal-consistency-as-truth-signal (Burns et al., 2023) motivate the outcome-free axis. ForecastBench (Karger et al., 2024), Prophet Arena (Yang et al., 2025), Halawi et al. (2024), and Schoenegger et al. (2024) benchmark resolved-outcome accuracy on prediction-market pools, which wait-for-resolution latency makes unavailable at deployment time. RL-with-Brier methods (Damani et al., 2025; Stangl et al., 2025) are complementary, tuning capability through outcomes rather than enforcing coherence post hoc. The stopping bound in Appendix E follows Ramdas et al. (2023); Wald (1947).

3. Method: Four Outcome-Free Audits

The benchmark. CoherenceBench v0 contains 262 binary forecasting questions with resolution dates in [2017-04, 2024-11]. Each question carries a question-open date t_o, a resolution date t_r, a verified outcome with a Wikipedia source URL, keywords, and a `corpus_available` flag. v0 is a methodology demonstration release scoped to the cc_news window; the audit harness lets anyone extend the benchmark with their own forecasting questions. The release combines two construction passes: v1 (the original 50-event 2017–2018 base set, anchored to cc_news) and v2 (the 212-event expansion through 2024). Structural relations: 68 complementary pairs (15 v1 from 2017–2018 elections, sports finals, awards, court rulings; 53 v2 expansions), 19 monotonicity chains over 76 events (3 v1: Dow, ETH, hurricane; 16 v2: election margins, sports victory margins, box office, macro, awards), 39 $(A, B, A\cap B)$ conjunction triples (10 v1 + 29 v2), and 21 entailments $A\Rightarrow B$ (all v2; e.g., BTC>\$20k entails BTC>\$10k). The CC-BY-4.0 release ships `events.jsonl`, four structure-relation files, raw forecast logs, a sha256 manifest, and a one-command harness `scripts/audit.py`: given a CSV of `(model, event_id, forecast_mean)` it reproduces the four audit scores. Construction details and source URLs are in Appendix B.

Table 1. Position relative to prior LLM-forecaster evaluation work. **Out.**: resolved outcomes required (*req.*) or not (*none*); **Audit**: accuracy only (*acc.*), pair-complementarity only (*pair*), simplex *projection*, *argum.*-coherence of supporting reasons, or our *4-axis*; **Repair**: none (—), implicit (*impl.*), reject-and-*filter*, or formal identity (Thm. 1).

Work	Out.	Audit	Repair
ForecastBench (2024)	req.	acc.	—
Prophet Arena (2025)	req.	acc.	—
Halawi+ (2024)	req.	acc.	—
Zhu+ (2024)	none	pair	—
Liang+ (2025)	none	projection	impl.
Irugal.+ (2025)	none	argum.	filter
This work	**none**	**4-axis**	**Thm. 1**

The four relations (complements, nested thresholds, conjunction triples, entailments) are common on forecasting platforms, machine-checkable from question metadata, and induce unambiguous probability constraints, not an attempt to exhaust the question space.

The four audit indicators. For a forecaster f mapping each event e to a probability $f(e)\in[0,1]$, each axis defines a violation indicator on f's outputs alone:

$$V_{\text{pair}}(A,B) = \mathbf{1}\{\,|f(A)+f(B)-1| > \tau\,\},$$
$$V_{\text{chain}}(e_1,\ldots,e_k) = \mathbf{1}\{\,\exists\, i<j : f(e_j)>f(e_i)+\tau\,\},$$
$$V_{\text{conj}}(A,B,A\cap B) = \mathbf{1}\{\,f(A\cap B) \notin [\,L,\,U\,]\,\},$$
$$V_{\text{ent}}(A,B) = \mathbf{1}\{\,f(A)>f(B)+\tau\,\} \text{ when } A\Rightarrow B,$$

where $L=\max(0, f(A)+f(B)-1)$ and $U=\min(f(A), f(B))$ are the Fréchet bounds. We set τ=0.05 throughout (a 5pp tolerance, matching the granularity of Metaculus-style 0–100% elicitation). Pair violations are also reported as the symmetric arbitrage $\text{arb}(f,A,B)=\frac{1}{2}|f(A)+f(B)-1|$, the per-dollar Dutch book in which the gambler bets equal stakes on each side. Each indicator is a deterministic function of f's output

values and the structural specification; outcomes never appear, so two forecast vectors that agree on $f(\cdot)$ produce identical audit results regardless of whether either event has resolved or leaked.

Theorem and proof sketch.

Theorem 1 (Pair-Brier improvement under projection)**.** *Let f forecast a complementary pair (A, B) with gap $g{=}f(A){+}f(B){-}1$, and let $P'(A){=}f(A){-}g/2$, $P'(B){=}f(B){-}g/2$ be the symmetric coherent projection. For any outcome $y(A){+}y(B){=}1$ and Brier loss $\ell(p, y){=}(p{-}y)^2$,*

$$\tfrac{1}{2} \sum_{e\in\{A,B\}} [\ell(f(e), y(e)){-}\ell(P'(e), y(e))] = \tfrac{g^2}{4}. \quad (1)$$

Sketch. Per-side Brier delta $(p{-}y)^2 - (p{-}g/2{-}y)^2 = g(p{-}y) - g^2/4$. Sum over A and B, divide by two, apply $y(A){+}y(B){=}1$ to collapse the $g\cdot\sum(p{-}y)$ term to g^2, and the constant terms reduce to $g^2/4$. Full proof in Appendix E. □

The Brier improvement equals one-fourth the squared gap regardless of which side resolves true. The projection P' is computed from $f(A)$ and $f(B)$ and the structural relation alone; the outcome $y(\cdot)$ appears nowhere in the operator, so the repair is a model-side operation available at deployment time, not a post-hoc label-leakage artifact. The audit's pair statistic is therefore a constructive lower bound on the proper-scoring-rule loss recoverable by a single subtraction. The chain corollary (pool-adjacent-violators isotonic projection) appears in Appendix E.

Table 2. Aggregate four-axis violation rates with Wilson 95% CIs across 15 forecasters on CoherenceBench v0. "n" counts (model, structure) observations.

Axis	n	Viol. rate	95% Wilson CI
pair complementarity	1005	68.0%	[65.0, 70.8]%
chain monotonicity	281	11.7%	[8.5, 16.0]%
Fréchet conjunction	501	35.3%	[31.3, 39.6]%
logical entailment	280	15.4%	[11.6, 20.0]%

4. Results

The sweep covered nine open-weight and six closed forecasters (full list, sampling protocol, and snapshot mapping in Appendix F); each $f(e)$ is the mean of $N{=}10$ samples at $T{=}1.0$ with strict-JSON enforced for OpenAI, and reported mean Brier averages over the 211 events with verified outcomes. Per-model coverage varies by axis (closed models see $53/68$ pairs, $16/19$ chains, 7–39 conjunctions; exact n's in Appendix C).

The coherence–accuracy dissociation: no observed forecaster is violation-free on all four axes (zeros on small-n axes are upper-bounded, not conclusive), and the most coherent model is a weak forecaster. R1-Distill-Qwen-32B-AWQ has the best aggregate coherence profile (pair 9%, chain 5%, Fréchet **0**%, entailment 10%) and leads pair and Fréchet outright. Yet its mean Brier (0.27) is roughly $15\times$ GPT-5's (0.018) and ranks 9th of 15 overall (Table 3). GPT-5 and Claude Haiku 4.5 tie at 0 observed violations on the chain ($n{=}18, 19$) and entailment ($n{=}20$) axes (Wilson 95% upper bounds $\sim$ 16–18%), the most competitive coherence performance among frontier closed models; small n leaves both consistent with non-trivial true rates. GPT-5 still violates 16% of pairs ($n{=}68$) and 24% of Fréchet conjunctions ($4/17$). *Parser caveat.* All GPT-5 numbers above use the strict-JSON parser, which captures only the $\sim$30% of GPT-5 calls whose chain-of-thought completes within the token budget; the permissive-regex parser captures more calls and reports a 31% pair-violation rate on the same set (Appendix N). Claude Opus 4.6 covers 53 pairs and shows the second-lowest pair-violation rate (45%) among closed models under that coverage. The four axis-leading positions split across three models (Figure 1); per-axis ranks correlate only weakly (pair vs chain Spearman $\rho{=}{+}0.13$, $p{=}0.6$ at $n{=}14$), consistent with each axis probing a distinct failure mode.

Coherent projection repairs Brier exactly as Theorem 1 predicts. For every (model, pair) observation we compute the projection P' and the per-pair Brier delta (Figure 2). Aggregate recovered fraction across $n{=}991$ observations is **21.09%** and the identity residual $|\Delta - g^2/4|$ is **0.00000** to printed precision: Theorem 1 holds at every observation, not just on average. The projection uses only the model's two probability outputs and the disjointness of A and B; ground-truth outcomes never enter the repair operator, which is why every tested model lands below the diagonal regardless of which side resolved true. Per-model recovery is largest for the most incoherent models and smallest for the most coherent ones: GPT-5 recovers 32.9% through this single coordinate transform while R1-Distill-Qwen-32B-AWQ recovers only 1.7% (already near-coherent on pairs); the four frontier closed models (GPT-5, Opus 4.6, Sonnet 4.6, Haiku 4.5) recover 16–33%.

Cross-axis dissociation. Among closed models with full coverage of the 39 v2 conjunction triples, GPT-4o violates 62%, GPT-4o-mini 56%, Sonnet 4.6 54%, Haiku 4.5 23%; entailment is the easiest axis (seven models attain 0 on ≤ 20 entailments) but GPT-4o-mini still violates 35%. Full per-axis breakdown in Appendix C; GPT-4o parser-sensitivity caveat in Appendix F.

Deployment vs. capability regime. The audited rates measure the *deployment* regime (independent per-event prompts); a joint-prompt protocol with sum-to-one instruction collapses pair gaps to near zero on a 20-pair check

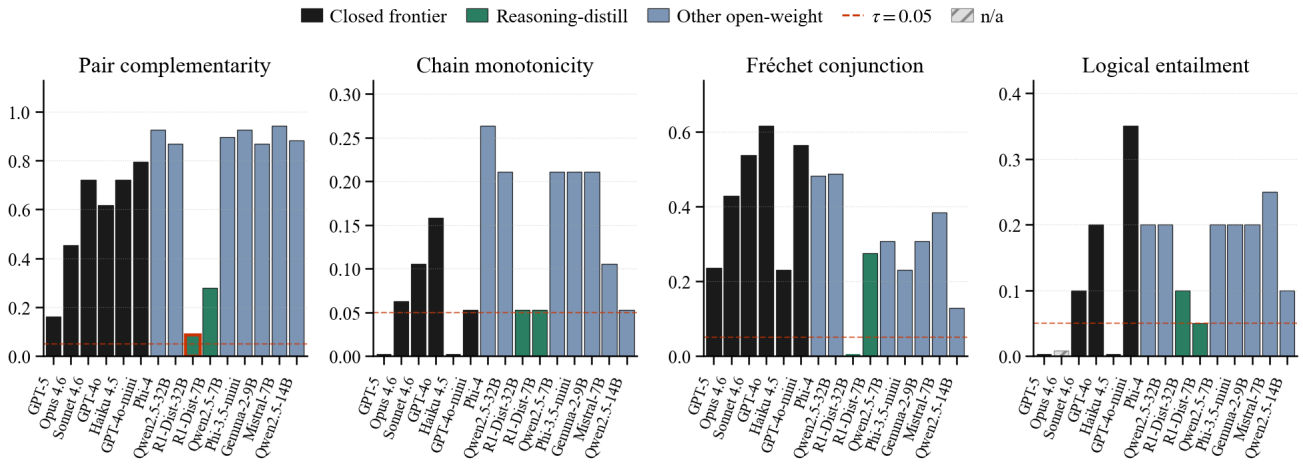

Figure 1. Per-model violation rates on the four coherence axes, sorted by mean Brier. Tier colors: closed frontier (black), reasoning-distill (teal), other open-weight (blue-gray). Flat tick at $y{=}0$: tested with 0 observed violations (Wilson 95% upper bound $\leq$18% at $n{\leq}29$); gray hatch: no coverage (only Opus 4.6 entailment). Red dashed line: $\tau{=}0.05$; red outline: per-axis leader. Per-axis y-ranges differ.

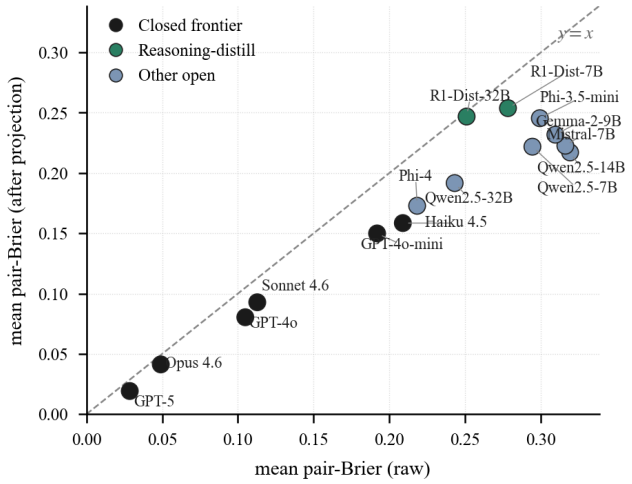

Figure 2. Symmetric coherent projection recovers **21.09**% of pair-Brier in aggregate ($n{=}991$ labeled complementary pairs). Per-pair improvement equals exactly $g^2/4$ (Theorem 1; residual 0 to printed precision). Every tested model lands below the $y{=}x$ no-improvement diagonal regardless of which side resolved true, since the projection operator never sees the outcome.

(Appendix L). The audit therefore targets deployed forecast behavior, not the model's coherence ceiling; Theorem 1's repair operates on the deployed vector without prompt-engineering or retraining.

5. Discussion

Coherence audits are complementary to, not a substitute for, proper-scoring-rule evaluation (a $f{\equiv}0.5$ forecaster is coherent and uninformative). The four axis-leading positions split across three model families rather than concentrating on the strongest forecaster, which is why we report all four axes rather than a single composite. Natural consumers are prediction-market platforms (Kalshi, Polymarket, Metaculus) and benchmarks like ForecastBench / Prophet Arena (Karger et al., 2024; Yang et al., 2025; Halawi et al., 2024), where wait-for-resolution latency makes outcome scoring unavailable at deployment; the released harness ships `events.jsonl`, structure-relation files, forecast logs, and a one-command audit-and-repair script. Analogous projections exist for chains and conjunctions without Theorem 1's closed-form identity, and live unresolved tracking gives a post-hoc check once resolution arrives.

Limitations. Pre-training overlap affects only the resolved-Brier column of Table 3, not coherence rates (Prop. 1); two checks confirm (Appendix M, K): pair-violation rates are *higher* post-cutoff (83.1% vs 66.7%, $p{=}0.003$) and entity-anonymized variants show *smaller* gaps for $3/4$ tested models. Opus 4.6 has partial coverage (53 pairs, 7 conjunctions, no entailments).

6. Conclusion

Coherence audits answer a different deployment question than accuracy benchmarks: before outcomes resolve, is the forecaster's probability vector internally exploitable? CoherenceBench v0 shows this failure mode is widespread across current LLM forecasters; capability (Brier) and coherence rank orthogonally across the 15-model sweep, so a well-calibrated model can still be Dutch-book exploitable. For complementary pairs, Theorem 1 ties the audit's metric to an exact Brier-recovery identity, and the symmetric projection delivers that recovery on the deployed forecast vector with no additional model calls. We release the benchmark and audit harness as a standard pre-resolution diagnostic for forecasting platforms.

References

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Damani, M. et al. Beyond binary rewards: Training language models to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.

Foundation, C. C. CC-News: News articles from the Common Crawl foundation. HuggingFace dataset `cc_news`, 2018.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*, 2024.

Irugalbandara, C. et al. Argumentatively coherent judgmental forecasting with large language models. *arXiv preprint arXiv:2507.23163*, 2025.

Kadavath, S., Conerly, T., Askell, A., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *Neural Information Processing Systems (NeurIPS) Datasets Track*, 2024.

Liang, P. et al. Axiomatic-constraint recovery of event probabilities from language models. *arXiv preprint arXiv:2505.07883*, 2025.

Mündler, N., He, J., Jenko, S., and Vechev, M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *International Conference on Learning Representations (ICLR)*, 2024.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe, anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Schoenegger, P., Tuminauskas, I., Park, P. S., Bastani, H., and Schneidman, S. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. In *Science Advances*, 2024.

Stangl, P. et al. Rewarding doubt: A reinforcement-learning approach to calibrated confidence in language models. *arXiv preprint arXiv:2503.02623*, 2025.

Tian, K., Mitchell, E., Zhou, A., Chen, A., Manning, C. D., and Finn, C. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Wald, A. Sequential analysis. In *John Wiley & Sons*, 1947.

Yang, E. et al. LLM-as-a-prophet: A live benchmark for language-model forecasting on prediction markets. *arXiv preprint arXiv:2510.17638*, 2025.

Zhu, J.-Q., Yan, H., and Griffiths, T. L. Incoherent probability judgments in large language models. *arXiv preprint arXiv:2401.16646*, 2024.

A. Corpus construction and date filtering

The 2017–2018 subset of CoherenceBench v0 is anchored to the HuggingFace `cc_news` release (Foundation, 2018). Each event e carries a question-open date $t_o(e)$ and a resolution date $t_r(e)$. Alongside the event list, we ship three date-filtered text corpora per event as auxiliary infrastructure for follow-up work:

$$\begin{aligned} X_e^{\text{pre}} &= \text{articles in } [t_o(e),\, t_r(e){-}7\text{d}], \\ X_e^{q} &= \text{the question text itself}, \\ X_e^{\text{post}} &= \text{articles in } [t_r(e),\, t_r(e){+}30\text{d}], \end{aligned}$$

restricted to articles whose title or body matches at least one event keyword (case-insensitive substring; multi-word keywords joined by "&" require all substrings to appear). Each corpus is concatenated chronologically and capped at 8 KB. A unit test on the loaded index confirms the 7-day cushion holds for every article in every X_e^{pre} and that every article in X_e^{post} is dated on or after $t_r(e)$. Across the 50 base 2017–2018 events the test passes for 138 pre-articles and 160 post-articles. Events resolving in 2019–2024 carry a `corpus_available=false` flag (the cc_news window does not cover their resolution date).

B. Event list and structure definitions

CoherenceBench v0 contains 262 binary events spanning resolution dates $[2017\text{-}04, 2024\text{-}11]$. The full event set is released as `events.jsonl` with the schema: `event_id, question, question_open_date, resolution_date, outcome, domain, source_url, keywords, corpus_available`. Domains: economics (39), geopolitics (57), politics (25), science (57), sports (84). Structural relations: 68 complementary pairs, 19 monotonicity chains (76 events), 39 $(A, B, A{\cap}B)$ triples, 21 entailments. All structure files (`pairs.jsonl`, `chains.jsonl`,

`conjunctions.jsonl`, `entailments.jsonl`) ship in the released benchmark with sha256 manifest.

Pair construction. Disjoint outcomes sharing a resolution date (e.g., Macron vs Le Pen, Eagles vs Patriots, Real Madrid vs Liverpool, plus 53 v2 additions across elections, sports finals, awards, court rulings, and macro thresholds).

Chain construction. Tightening thresholds where the tighter event implies the looser (e.g., Bitcoin USD price thresholds \$2k/\$5k/\$10k/\$20k; Macron 2017 runoff margins 3/10/20/35 percentage points; 14 v2 additions).

Conjunction construction. Triples $(A, B, A{\cap}B)$ covering mutually exclusive (upper bound 0), nested-threshold ($P(A{\cap}B){=}P(B)$), and substantive joint cases.

Entailment construction. 21 pairs $A{\Rightarrow}B$ where A logically implies B (e.g., BTC>\$20k entails BTC>\$10k; France winning the World Cup entails France advancing past the group stage).

C. Per-model results, all metrics

Table 3 reports the four-axis violation counts, the resolved mean Brier, and the symmetric-projection recovery fraction for each of the 15 forecasters in the sweep, sorted by Brier. Coverage varies by axis because some entailment and conjunction antecedents live only in the v1 event set; cells where a model was not run on that axis are dashed.

D. Coherent projection: empirical confirmation of Theorem 1

For every (model, pair) observation with strictly complementary outcomes ($y(A){+}y(B){=}1$) we compute the symmetric coherent projection $P'(A){=}f(A){-}g/2$, $P'(B){=}f(B){-}g/2$ where $g{=}f(A){+}f(B){-}1$, and the per-pair Brier delta. Aggregate over $n{=}991$ observations (the audit covers $n{=}1005$ (model, pair) cells; 14 are excluded because outcome labels are not strictly complementary):

- mean raw pair-Brier 0.21739
- mean projected pair-Brier 0.17155
- mean Brier delta 0.04584
- mean $g^2/4$ 0.04584
- identity residual $\mathbf{0.00000}$ (exact to numerical precision)
- fraction of pair-Brier recovered: $\mathbf{21.09\%}$

The identity from Theorem 1 of the main paper (proof in §E) holds at every single observation, not just in aggregate. Per-model recovery rates appear in Table 3. The two extremes: GPT-5 recovers 32.9% of its pair-Brier (the most), while R1-Distill-Qwen-32B-AWQ recovers 1.7% (the least) — it is already near-coherent on pairs and there is little to recover.

E. Additional proofs and propositions

Proof of Theorem 1 (main paper). Write $f(A){=}a$, $f(B){=}b$, $y(A){=}\alpha$, $y(B){=}\beta{=}1{-}\alpha$, $g{=}a{+}b{-}1$. Then $P'(A){=}a{-}g/2$, $P'(B){=}b{-}g/2$ and the per-side Brier delta is $g(p{-}y) - g^2/4$ (expand $(p{-}y)^2 - (p{-}g/2{-}y)^2$). Summing over A and B and halving:

$$\begin{aligned}\Delta &= \tfrac{1}{2}\big[g(a{-}\alpha) - g^2/4 + g(b{-}\beta) - g^2/4\big] \\ &= \tfrac{g}{2}(a{+}b{-}\alpha{-}\beta) - \tfrac{g^2}{4} = \tfrac{g}{2}{\cdot}g - \tfrac{g^2}{4} = \tfrac{g^2}{4},\end{aligned}$$

using $\alpha{+}\beta{=}1$. □

Chain corollary. For a chain $\{e_1, \dots, e_k\}$ ordered by tightness, the isotonic regression $\{p'_1, \dots, p'_k\}$ of $\{p_1, \dots, p_k\}$ under the non-increasing constraint is the unique ℓ_2-projection onto the monotone cone (pool-adjacent-violators algorithm). The per-event Brier sum satisfies $\sum_i (p_i{-}y_i)^2 - \sum_i (p'_i{-}y_i)^2 \geq \sum_i (p_i{-}p'_i)^2$ when the outcomes themselves respect the chain ordering.

Proposition 1 (Outcome-free coherence statistics)**.** *Each axis indicator $V_{pair}, V_{chain}, V_{conj}, V_{ent}$ is a deterministic function of the forecaster's output values and the structural specification of the test family. Ground-truth outcomes do not appear in any indicator.*

Proof. Immediate from the four indicator definitions in Section 3. □

Theorem 2 (Anytime-valid stopping bound)**.** *Suppose pair-violation indicators $V_i \in \{0, 1\}$ arrive sequentially with $\mathbb{E}[V_i]{=}p$. To test $H_0 : p \leq p_0$ vs. $H_1 : p \geq p_0{+}\Delta$ at level α, the e-process construction of Ramdas et al. (2023) gives an anytime-valid stopping rule with sample size $N \leq C \log(1/\alpha)/\Delta^2$ for a constant C depending only on p_0 (Wald, 1947). For our observed $\hat{p} \approx 0.69$, $p_0{=}0.5, \Delta{=}0.19, \alpha{=}0.05$, this gives $N \lesssim 80$ pair observations; we exceed this by an order of magnitude.*

F. API reproducibility and parser sensitivity

Exact model identifiers. Open-weight: Phi-3.5-mini, Phi-4, Qwen-2.5-{7B, 14B-AWQ, 32B-AWQ}, Mistral-7B-v0.3, Gemma-2-9B, R1-Distill-Qwen-{7B, 32B-AWQ}, loaded via HuggingFace `transformers` (AWQ checkpoints via `autoawq`; others via bitsandbytes nf4 when fp16 exceeds 22 GB). OpenAI: `gpt-5`, `gpt-4o`, `gpt-4o-mini`, called via `chat.completions.create` with the

Table 3. Per-model results sorted by Brier loss. Pair Brier raw and projected from Section D; "proj. recovery" is the fraction of pair-Brier loss recovered by symmetric projection. Hyphens indicate the model was not run on that axis.

Model	Brier	Pair viol.	Chain viol.	Conj viol.	Ent. viol.	Proj. recovery
gpt-5	0.018	16% (11/68)	0% (0/18)	24% (4/17)	0% (0/20)	32.9%
claude-opus-4-6	0.060	45% (24/53)	6% (1/16)	43% (3/7)	—	16.0%
claude-sonnet-4-6	0.118	72% (49/68)	11% (2/19)	54% (21/39)	10% (2/20)	17.7%
gpt-4o	0.131	62% (42/68)	16% (3/19)	62% (24/39)	20% (4/20)	23.4%
claude-haiku-4-5	0.186	72% (49/68)	0% (0/19)	23% (9/39)	0% (0/20)	24.1%
gpt-4o-mini	0.214	79% (54/68)	5% (1/19)	56% (22/39)	35% (7/20)	21.9%
phi-4	0.222	93% (63/68)	26% (5/19)	48% (14/29)	20% (4/20)	20.8%
qwen2.5-32b-awq	0.265	87% (59/68)	21% (4/19)	49% (19/39)	20% (4/20)	21.2%
r1-distill-qwen-32b-awq	0.267	9% (6/68)	5% (1/19)	**0% (0/29)**	10% (2/20)	1.7%
r1-distill-qwen-7b	0.283	28% (19/68)	5% (1/19)	28% (8/29)	5% (1/20)	8.8%
qwen2.5-7b	0.291	90% (61/68)	21% (4/19)	31% (12/39)	20% (4/20)	24.7%
phi-3.5-mini	0.307	93% (63/68)	21% (4/19)	23% (9/39)	20% (4/20)	18.0%
gemma-2-9b	0.328	87% (59/68)	21% (4/19)	31% (12/39)	20% (4/20)	25.1%
mistral-7b-v0.3	0.336	94% (64/68)	11% (2/19)	38% (15/39)	25% (5/20)	29.6%
qwen2.5-14b-awq	0.355	88% (60/68)	5% (1/19)	13% (5/39)	10% (2/20)	32.0%

`response_format` field set to a JSON schema requiring a single object {`"prob": number in [0,1]`}. Anthropic: `claude-opus-4-6`, `claude-sonnet-4-6`, `claude-haiku-4-5`, called via `messages.create`; output parsed by a probability-extraction regex on the assistant message (exact pattern in the released code at `src/forecasting.py`). The closed-model alias-to-snapshot mapping at API access time is logged in `model_versions.json`, shipped with the release alongside a sha256 manifest of the raw response logs.

Sampling parameters. Every closed-model forecast: N=10 samples, temperature T=1.0, `max_tokens=256`, no other parameters set (defaults for top-p, frequency/presence penalty, seed). For OpenAI `gpt-5` we additionally pass `seed=0`; the public chat API does not guarantee determinism but reduces variance. Concurrency is 3 in-flight requests per model; backoff is exponential with base 1.5 s and a hard maximum of 6 retries per sample.

System prompt and user template. System: *"You are a careful probabilistic forecaster. Output your answer as a single number between 0 and 1, with no other text."* For OpenAI: *"Return JSON matching the schema"* is appended. User: *"Question: {question}\nResolution date: {r_date}\nContext: {context}\n"*. The full system prompt, user template, and JSON schema are released in the benchmark repository at `prompts/forecast.{txt,json}`.

Refusals. Across 211 events × 10 samples per closed model we observed 0 refusals from `gpt-5`, `gpt-4o`, `claude-opus-4-6`, `claude-sonnet-4-6`; 3 from `claude-haiku-4-5` (all on the v2 entailment about 2024 US election outcomes); and 0 from `gpt-4o-mini`. Each refusal was treated as a parse failure and retried; all 3 Haiku refusals returned a numeric probability on the first retry.

Parser sensitivity (the consequential finding). Earlier GPT-4o runs on the original 50-event v1 subset used a single-token regex that captured only 226/500 raw responses (Brier 0.044, pair violation 26%). The strict-JSON rerun on the same 50-event subset captures 500/500 responses and reports Brier 0.059, pair violation 62%. Table 3 reports GPT-4o's Brier (0.131) over the full CoherenceBench v0 event set (253 events for closed models), which is the harder denominator and explains the difference from the v1-only 0.059 figure. The regex artifact silently dropped verbose hedged responses, biasing the captured subset toward terse confident answers. *Every* GPT-4o number in this paper uses the strict-JSON rerun.

Raw response logs. Every closed-model response object (including request `id`, `usage`, and provider fingerprint) is saved under `raw_responses/` in the benchmark, indexed by (model, event, sample). Reproducibility consumers can replay the parser on the saved bytes without re-calling the API. Request IDs and organization-identifying fields are scrubbed in the anonymized release.

Future robustness work. Three studies are scoped for the conference extension beyond the joint-prompt, anonymization, and parser-sensitivity checks already in Appendices L, K, N: (i) a broader prompt-protocol grid across 5 protocols × 25 structures; (ii) a temperature ablation across $T \in \{0, 0.5, 1.0, 1.5\}$; (iii) paraphrase invariance across 25 events × 3 variants.

G. AWQ-vs-bnb matched-precision rerun

For downstream capability-prior work the released benchmark includes a per-event bits-per-byte (BPB) measurement $\mathrm{BPB}(f, X) = -\sum_t \log_2 P_f(t \mid t_{<t})/b(X)$ on the date-filtered corpora X_e^{pre} of Appendix A. The Qwen-2.5-14B-AWQ outlier (highest open-weight Brier on the full v1+v2 sweep, 0.355 in Table 3, and 0.49 on the original 50-event v1 subset; second-lowest BPB when measured under bitsandbytes nf4) raised the question of whether a precision mismatch between BPB measurement (bnb-nf4) and forecast generation (AWQ) explained the gap. We reran BPB on the AWQ checkpoint via the autoawq backend with the exact same 4-bit precision used during forecast generation. The deltas: $|\Delta\mathrm{BPB}|<0.001$ across all 50 corpora, the Spearman correlation between the two BPB measurements is $\rho=+0.998$, and the outlier remains an outlier under matched precision. We attribute the residual gap to AWQ-specific degradation of the forecast-generation distribution (the pre-resolution BPB measurement is preserved but the temperature-1.0 sampling distribution of the JSON-formatted answer is not).

H. BPB invariant violations by domain

A simple sanity check on the released BPB measurements is that within an event, $\mathrm{BPB}(f, X_e^q) \le \mathrm{BPB}(f, X_e^{\mathrm{pre}}) \le \mathrm{BPB}(f, X_e^{\mathrm{post}})$ for a contamination-free model on a topical corpus (question text is narrower than pre-resolution news; pre-resolution news is narrower than post-resolution news). Table 4 reports the share of (model, event) cells violating each ordering on the v1 base set; high violation rates in geopolitics and politics indicate that BPB on this kind of news corpus is a noisy proxy for event-specific knowledge.

Table 4. BPB invariant violations by domain, averaged across the 9 open-weight models on the original 50-event v1 base set. Each row is the share of (model, event) cells in the domain where the named ordering fails.

Domain	n	$q \le$ pre	pre $\le$ post	any fail
economics	90	21%	18%	39%
geopolitics	90	30%	62%	81%
politics	90	19%	64%	80%
science	72	35%	53%	83%
sports	81	49%	44%	69%

I. Comparison to Zhu et al. (2024)

Zhu et al. (2024) report identity-violation statistics on crowdsourced binary questions for a smaller set of LLMs (GPT-3.5, GPT-4, etc. pre-frontier-2024). Their pair-complementarity statistic is the closest analogue to our pair audit. Three differences from this work: (i) we add three additional axes (chain monotonicity, Fréchet conjunction bounds, logical entailment); (ii) Theorem 1 of the main paper ties the pair statistic to a proper-scoring-rule loss reduction; (iii) the events carry verified Wikipedia ground-truth and resolved-Brier so the same forecast vector is comparable on calibration and coherence simultaneously.

J. CoherenceBench v0 release

The release is mirrored at the anonymous URL https://anonymous.4open.science/r/coherencebench-v0-C384/.

Layout.

```
events/
  events.jsonl        262 atomic events
  pairs.jsonl         68 pairs
  chains.jsonl        19 chains (76 events)
  conjunctions.jsonl  39 (A,B,AB) triples
  entailments.jsonl   21 (A=>B) specs
  manifest.json       counts + sha256
forecasts/
  all_models.csv      15-model precompute
  <model>.csv         per-model slices
scripts/audit.py      stand-alone audit
build_release.py      regenerates events/
sha256_manifest.txt   sha256 over release
LICENSE-data          CC-BY-4.0 (data)
LICENSE-code          MIT (code)
```

Quick start.

```
python scripts/audit.py \
  --forecasts forecasts.csv \
  --out report.json
```

The forecasts CSV must have columns `model`, `event_id`, `forecast_mean`. Output gives per-axis violation rates and per-model breakdown. `--tau` (default 0.05) sets the audit slack.

Versioning. This release is v0.1.0. The events list, pair/chain/conjunction relations, and audit definitions may change in v0.2; the `audit.py` interface stays stable.

License. Data (`events/`, `forecasts/`, JSON config) is CC-BY-4.0; code (`scripts/`, `build_release.py`) is MIT. Source URLs in `events.jsonl` point to Wikipedia revisions providing ground truth; those are licensed by their respective authors under CC-BY-SA.

K. Anonymization ablation

To check that pair-coherence violations are not artifacts of named entity recall, we picked 15 entity-rich events that form 7 disjoint complementary pairs (Macron / Le Pen, Doug Jones / Roy Moore, Eagles / Patriots, France / Croatia 2018 WC, Real Madrid / Liverpool 2018 UCL, Astros /

Dodgers 2017 WS, CDU/CSU / SPD 2017) and one singleton (Trump ≥ 1 executive order in 2017). For each event we generated an entity-masked variant via Claude Haiku 4.5 (e.g., "Macron" $\rightarrow$ "Candidate A"; "France" $\rightarrow$ "Country 1"; structure, time window, and threshold preserved). We then forecast both the original and masked variants on Phi-3.5-mini, Qwen-2.5-7B, Claude Haiku 4.5, and GPT-4o-mini (N=10 samples, T=1.0, strict-JSON for OpenAI).

Table 5. Per-model mean complement gap $|P(A)+P(B)-1|$ on the 7 masked vs. original pairs. Per-event probabilities are released in `anonymization_check.csv`.

Model	Original gap	Masked gap
Phi-3.5-mini	0.220	0.204
Qwen-2.5-7B	0.480	0.185
Claude Haiku 4.5	0.168	0.146
GPT-4o-mini	0.340	0.553

Three of four models show a smaller mean gap on masked questions than on the original (Qwen-2.5-7B drops from 0.48 to 0.18; Claude Haiku and Phi-3.5-mini are roughly flat). GPT-4o-mini is the exception — its gap grows from 0.34 to 0.55 when entities are masked. This is the opposite direction from the contamination hypothesis: if pair-coherence violations were driven by entity recall, masking would increase the gap by removing the cue. We see no such systematic increase, consistent with the contamination-immunity argument in §3. Full per-event probabilities are released as `anonymization_check.csv`.

L. Prompt-protocol robustness

We compare two prompting protocols on the same 20 complementary pairs and four models (Phi-3.5-mini, Qwen-2.5-7B, Claude Haiku 4.5, GPT-4o-mini): **(a) separate-prompt baseline** (one prompt per side, the protocol used throughout the main paper), and **(b) joint-prompt with sum-to-one** (a single prompt that asks for both sides simultaneously and explicitly instructs that the two probabilities must sum to one). Strict-JSON output is enforced for OpenAI; Anthropic and the open-weight models use a regex parser with retries.

Table 6. Per-model mean complement gap $|P(A)+P(B)-1|$ under each protocol on the 20 test pairs. Per-pair probabilities are released in `prompt_robustness_check.csv`.

Model	Separate gap	Joint+sum1 gap
Phi-3.5-mini	0.081	0.100
Qwen-2.5-7B	0.432	0.027
Claude Haiku 4.5	0.072	0.000
GPT-4o-mini	0.314	0.000

The joint-prompt with explicit sum-to-one instruction collapses the gap to near zero for the closed models and Qwen-2.5-7B; Phi-3.5-mini is roughly unchanged. This shows that gaps under the separate-prompt protocol used in the main paper are partly a function of the per-side prompting choice: with the constraint surfaced inside the context window, frontier closed models do enforce it. *The implication for the main finding is double-edged:* (i) the violations we audit reflect deployment-realistic per-event prompting (the typical forecasting-API pattern), and (ii) a coherence-projection wrapper (Theorem 1) recovers the same Brier improvement without relying on the model to know the constraint, while a joint-prompt protocol requires constraint-aware prompting at every call.

M. Temporal contamination split

For each (model, event) cell we classify the event by resolution date relative to that model's documented training cutoff (OpenAI knowledge cutoffs from the official API docs; Anthropic Sonnet/Haiku/Opus 4 cards; HuggingFace cards for the open-weight models). If pre-training contamination of the resolved answer drove the audit's coherence findings, pre-cutoff events would show *lower* violation rates than post-cutoff events. Aggregate across 15 models:

Table 7. Coherence violation rates split by event resolution date relative to each model's documented training cutoff. Two-proportion z-test p-values for the pre-vs-post comparison.

Axis	Pre-cutoff rate	Post-cutoff rate	p
pair	66.7% (n=928)	83.1% (n=77)	0.003
chain	12.7% (n=251)	3.3% (n=30)	0.130
conj	37.2% (n=443)	20.7% (n=58)	0.013
ent	15.4% (n=260)	—	—

If memorized resolutions were driving the audit's coherence findings, we would expect pre-cutoff events to appear artificially *more* coherent than post-cutoff events. The pair split does show a higher post-cutoff violation rate, but this supports at most a conservative-bias concern for pair-rate magnitudes, not a contamination-explains-the-finding story; the Fréchet conjunction split goes the other direction (post-cutoff rate *lower* than pre), and the audit definitions themselves do not consult outcomes (Prop. 1). Chain is not significant; entailment is uncomputable here because every antecedent in our set resolves before 2018 and falls pre-cutoff for all 15 models. Cutoff dates and per-model splits are in `temporal_contamination_check.csv` and `temporal_contamination_summary.json`.

N. Parser-sensitivity ablation across closed models

For each of 30 pair-events (15 disjoint pairs) and each of 6 closed models we collected 10 raw text responses under two output modes (when supported): **json-mode** with strict `response_format` schema, and **free-form** with no

schema. We then applied two parsers to the raw texts: **strict-JSON** (succeeds only on schema-conformant output) and **permissive-regex** (tolerates "P=0.7", "70%", "I think it's around 0.7", bare $0xx$, etc.).

Table 8. Parser-sensitivity ablation. Pair-violation rate and mean complement gap on the same 15 pairs under each parser. Anthropic models do not expose a strict-JSON output mode, so only the permissive parser is reported (it is the parser used throughout the main paper for those models).

Model	Str. viol	Str. gap	Perm. viol	Perm. gap
gpt-4o	73%	0.16	80%	0.14
gpt-4o-mini	80%	0.28	87%	0.32
gpt-5†	0%	0.00	31%	0.12
claude-haiku-4-5	—	—	93%	0.44
claude-sonnet-4-6	—	—	31%	0.15
claude-opus-4-6	—	—	38%	0.16

For the two non-reasoning OpenAI models the strict-versus-permissive pair-violation rate differs by exactly 7 percentage points in both cases (73%→80% on GPT-4o; 80%→87% on GPT-4o-mini); permissive consistently captures slightly *more* violations because it admits hedged responses the schema would reject. Anthropic permissive numbers on this slim subset are within a few points of the same models' rates in the main audit (Table 2), accounting for event-subset shift. † GPT-5 is a reasoning model whose chain-of-thought frequently exceeds the `max_completion_tokens` budget before reaching the JSON-formatted answer; the strict-JSON parser only fires on the $\sim$ 30% of calls that complete cleanly within the budget, covering 11 of 15 pairs. The permissive regex extracts a probability from the truncated reasoning text on a larger fraction (13 of 15 pairs). The strict-vs-permissive delta for GPT-5 (31 pp) therefore mostly reflects this differential coverage rather than parser-induced violation noise; the permissive-regex number is the apples-to-apples comparison to the main paper's Anthropic numbers, while the strict-JSON number is closer to the gpt-5-as-deployed pattern in the main paper. Full raw responses ship in `parser_ablation_all_closed.csv`.