Outcome-Free Arbitrage Audits and Coherence Repairs for LLM Forecasters

Juliana Li [1] **Diya Sreedhar** [1]

Abstract

LLM forecasters now rival crowd aggregates on resolved-outcome benchmarks, but they are deployed on questions that have not yet resolved, where no outcome is available to score them. The same probabilities can still be checked for coherence: ask a model for $P(A)$ and for its complement $P(A^c)$, and the two often fail to sum to one. In de Finetti's sense, coherence is a no-arbitrage condition, so an incoherent forecast vector is a self-inconsistent price system, detectable and repairable ahead of resolution. Reading each forecast as the price of a \$1 contract, we audit four classes of static arbitrage: complementary pairs, monotonicity chains, Fréchet conjunction bounds, and logical entailment. Across 15 forecasters on **CoherenceBench v0** (262 events), $\mathbf{68.8}\%$ of complementary pairs are incoherent and the mean implied arbitrage is $\$0.143$ per \$1 of gross notional. A single closed-form step, the symmetric projection onto the coherent set, certifies a per-pair Brier reduction of $g^2/4$ for gap g before any outcome is consulted (Theorem 1, the square of the per-dollar arbitrage); on the resolved benchmark this is $\mathbf{20.37}\%$ of pair-Brier, with no extra model calls. Accuracy does not reliably predict coherence (Spearman $\rho{=}{+}0.15$, 95% CI $[-0.42, +0.65]$, $n{=}15$): accurate models still quote arbitrage on most pairs, so coherence is a diagnostic in its own right. We release the benchmark and a one-command audit.

1. Introduction

LLM forecasters are increasingly run on questions that resolve months or years out, and on resolved-outcome benchmarks such as ForecastBench (Karger et al., 2024) and Prophet Arena (Yang et al., 2025) the best models rival crowd aggregates. But proper-scoring-rule loss is available only after resolution, so it is silent exactly when the deployment decision is made.

A different check needs no outcomes. Ask the model for $P(A)$ and for the complement $P(A^c)$: coherence requires $P(A){+}P(A^c){=}1$, so any gap $|P(A){+}P(A^c){-}1|$ is a Dutch book that pays regardless of how the event resolves, detectable when forecasts are collected. The two forecasts are paired outputs of one model on one prompt template and date, not independent draws. This self-disclosure keeps the audit outcome-free and unaffected by pre-training leakage: a model that has memorized a two-candidate or two-team outcome still fails the audit if its probabilities for the two sides do not sum to one.

Concretely, a model that puts 0.63 on an event and 0.52 on its negation prices the pair at 1.15: it is overpriced by $g{=}0.15$ across its two \$1 legs, i.e. 7.5 cents of arbitrage per \$1 gross notional. Symmetric repricing to 0.555 and 0.445 restores coherence and lowers the pair's mean Brier loss by $g^2/4{=}0.0056$, whichever way the event resolves. This illustrates the paper's premise: incoherence is visible before any outcome, and the repair has a closed-form score gain.

In deployment the audit acts as a static-arbitrage check in a pricing library: it flags inconsistencies that should not pass downstream, even where no label exists to score accuracy.

Coherence as no-arbitrage. de Finetti's coherence, meaning that no system of bets yields a sure loss, is the subjective-probability form of no-arbitrage (de Finetti, 1937; Seidenfeld et al., 2010; Nau, 2001), whose market-prices form is the Fundamental Theorem of Asset Pricing (Harrison & Kreps, 1979; Harrison & Pliska, 1981; Delbaen & Schachermayer, 2006): no arbitrage holds if and only if a consistent pricing measure exists. A coherent forecast vector is thus an arbitrage-free price system, and the coherent projection a minimal-ℓ_2 reprice onto the no-arbitrage manifold. Predd et al. (2009) show that any incoherent forecast is Brier-dominated by its coherent projection; Theorem 1 is the closed-form companion giving the exact amount.

Related work. Prior work tests LLM probabilistic coherence, including complement-parity violations (Zhu et al.,

[1]Harvard University, Cambridge, MA, USA. Correspondence to: Juliana Li <julianali@college.harvard.edu>.

Forecasting as a New Frontier of Intelligence Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

2024), simplex projection without a scoring-rule identity (Liang et al., 2025), and argumentative rather than numerical coherence (Irugalbandara et al., 2025), but none gives the no-arbitrage framing or the closed-form Brier recovery we derive. Resolved-outcome benchmarks (Karger et al., 2024; Yang et al., 2025; Halawi et al., 2024; Schoenegger et al., 2024) and RL-with-Brier methods (Damani et al., 2025; Stangl et al., 2025) evaluate or optimize accuracy after resolution; our audit is complementary, applying before outcomes. We also build on calibration (Kadavath et al., 2022; Tian et al., 2023) and self-consistency (Elazar et al., 2021; Mündler et al., 2024; Burns et al., 2023) work, which evaluates models against their own outputs.

Contributions. We give **(i)** a no-arbitrage reading of probabilistic coherence that audits any forecaster without outcomes (Section 2); **(ii)** the complement projection identity (Theorem 1): the coherent projection recovers exactly $g^2/4$ of pair-Brier, outcome-independent and in closed form; **(iii)** results on 15 forecasters showing widespread incoherence that accuracy does not predict (Section 3); and **(iv)** CoherenceBench v0 and a one-command audit.[1]

2. Method

Setup. Fix a finite set of events and a forecaster f, and read each value $f(e)\in[0,1]$ as the price of a contract paying \$1 if e occurs. de Finetti coherence is the absence of a static arbitrage (a zero-cost portfolio of these contracts with payoff ≥ 0 in every state and > 0 in some state; equivalently, positive cash up front with nonnegative payoff) over the priced events. We do not test the full no-arbitrage condition, which ranges over all portfolios; we test four machine-checkable classes, each pinning down a sharp relation among forecasts. Each indicator depends on f's outputs alone, so the audit consults no outcome and is unaffected by whether an event has resolved or leaked (Prop. 1).

The benchmark. CoherenceBench v0 is a released benchmark of 262 binary forecasting questions (resolution dates $[2017\text{-}04, 2024\text{-}11]$, five domains), each with a Wikipedia-verified outcome used only as a downstream cross-check. Four machine-checkable relations induce the constraints we test: 68 complementary pairs, 19 monotonicity chains (76 events), 39 $(A, B, A\cap B)$ conjunction triples, and 21 entailments $A\Rightarrow B$; a one-command harness reproduces the scores and repair from a forecast CSV (Section F).

The four indicators. For a forecaster f and tolerance τ (τ=0.05 throughout, a 5pp slack matching 0–100% elicitation granularity):

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

with $L=\max(0, f(A)+f(B)-1)$, $U=\min(f(A), f(B))$ the Fréchet bounds. Each is a familiar pricing constraint: complement parity (the two sides sum to \$1, a box-spread), strike monotonicity (a tighter threshold prices no higher than a looser one, e.g. $P(\text{BTC}>\$20\text{k}) \leq P(\text{BTC}>\$10\text{k})$), the Fréchet bound (a joint price stays within the range its marginals imply), and dominance ($A\Rightarrow B$ forces $P(A)\leq P(B)$). The complement gap $g=f(A)+f(B)-1$ has a dollar reading:

$$\text{arb}(f,A,B) = \tfrac{1}{2}\,|f(A)+f(B)-1| = \tfrac{1}{2}|g|, \qquad (1)$$

the arbitrage the symmetric book locks in per \$1 of gross notional (selling \$1 of each leg nets $|g|$ on \$2 of contracts).

The coherent projection. Repairing a complementary pair is one symmetric step: split the gap evenly between the two sides so they sum to one. The repair does not predict which side resolves; it removes an inconsistency present in both possible worlds. Because exactly one side of the pair is true, the outcome-dependent terms cancel and a fixed gain remains, the same number whichever way the event goes; Theorem 1 gives it in closed form.

Theorem 1 (Complement projection identity)**.** *For a complementary pair (A,B) with gap $g=f(A)+f(B)-1$, let $P'(A)=f(A)-g/2$, $P'(B)=f(B)-g/2$ be the symmetric coherent projection (the minimal-ℓ_2 projection onto $\{f(A)+f(B)=1\}$). For any outcome $y(A)+y(B)=1$ and squared-error (Brier) loss $\ell(p,y)=(p-y)^2$, a strictly proper scoring rule (Gneiting & Raftery, 2007; Savage, 1971),*

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

The identity has three consequences (proof in Section A). First, the recovered loss is outcome-independent: y does not enter P', so the projection runs on the forecasts as soon as they are collected, well before any outcome. Second, the recovered loss equals the square of the per-dollar arbitrage in (1), so an arbitrage of \$0.10 per \$1 corresponds to 0.10^2=0.01 of recovered Brier. Third, it is the exact Brier recovery of the minimal-ℓ_2 projection; Predd et al. (2009) show that any coherent projection Brier-dominates the incoherent forecast without giving the amount; the chain analog is the pool-adjacent-violators isotonic projection (Section A).

At n=1020 the 68.8% rate is not plausibly a small-sample artifact, far exceeding the size a sequential test needs to reject a $\leq 50\%$ rate (Ramdas et al., 2023; Wald, 1947).

[1] Code and data: https://github.com/lijuliana/coherencebench-v0

Table 1. Frequency of incoherence by class, with Wilson 95% CIs, across 15 forecasters on CoherenceBench v0. "n" counts (forecaster, structure) observations.

Constraint class	n	Violation rate	95% CI
complement parity	1020	68.8%	[65.9, 71.6]%
strike monotonicity	285	10.9%	[7.8, 15.0]%
Fréchet conjunction	510	35.1%	[31.1, 39.3]%
dominance (entailment)	287	15.7%	[11.9, 20.3]%

3. Results

The sweep covers nine open-weight and six closed forecasters (full protocol and snapshot mapping in Section E); each $f(e)$ is the mean of $N{=}10$ samples at $T{=}1.0$. We report how much incoherence the models carry, how much loss the projection recovers, and how coherence relates to accuracy.

(a) The implied arbitrage is large. Aggregated over 1005 complementary pairs,[2] the mean pair is internally inconsistent by **\$0.143** per \$1 of gross notional (a coherence metric; Section 4), and **68.8**% of pairs exceed the $\tau{=}0.05$ tolerance (Figure 1). The spread is wide and does not track size or vintage: small open-weight models (Qwen2.5-14B, Mistral-7B) leave \$0.26–\$0.28 per pair, while GPT-5 and the reasoning-distilled R1-Distill-32B leave only \$0.027 and \$0.019 and exceed the tolerance on just 13% and 9% of pairs, the most coherent in this sweep. Incoherence is therefore the common case in this field: the one constraint a coherent model satisfies exactly is the one most forecasters violate on a majority of pairs, with only those two staying close.

Complement parity is the dominant failure mode; the other three classes expose model-specific weaknesses (Figure 2). No model is clean on every axis, and none leads on all four, which is why we report all four rather than a single composite.

(b) The projection recovers loss as Theorem 1 predicts. For every (model, pair) we apply the projection P' and measure the squared-error delta (Figure 3). Aggregate recovery over $n{=}1005$ pairs is **20.37%** of pair-Brier, and the identity residual $|\Delta - g^2/4|$ is **0.00000** to printed precision, so (2) holds at every individual pair. Because the residual is zero, the recovery is not a noisy average: each pair gives back precisely the $g^2/4$ it owes, certified from the forecasts the moment they are collected, so the 20.37% aggregate is exact and outcome-independent. Every model lands below the $y{=}x$ no-improvement diagonal regardless of which side resolved true, because the projection never sees the outcome. Per-model recovery ranges from 1.7% (R1-Distill-32B, already near-coherent) to 36.5%, and it requires no retraining and no extra model calls.

[2]The four-class audit (Table 1) scores all 68 pairs for each of the 15 models (1020 observations); the dollar and projection metrics use the 1005 pairs priced by every model. Per-model rates can differ by one pair between the two sets (GPT-5: $9/67{=}13\%$ pooled, $10/68{=}15\%$ on the full set).

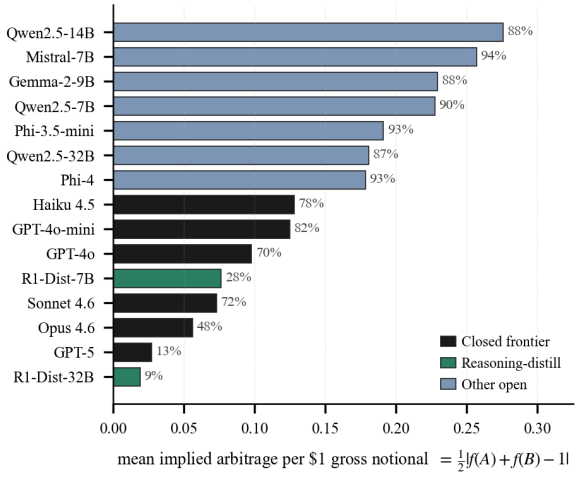

Figure 1. Mean implied arbitrage per \$1 of gross notional, $\frac{1}{2}|f(A)+f(B)-1|$, by forecaster (labels: % of pairs above the $\tau{=}0.05$ tolerance). Tier colors: closed frontier (black), reasoning-distill (teal), other open (blue-gray). GPT-5's low bar is confirmed by a full-capture re-run (Section D).

(c) Accuracy and coherence are not tightly coupled. Accuracy does not reliably predict coherence in this sample. With composite incoherence defined as the mean of a model's available class rates, its association with mean Brier is small and imprecise: Spearman $\rho{=}{+}0.15$ (Pearson $r{=}{+}0.30$), bootstrap 95% CI $[-0.42, +0.65]$, permutation $p{=}0.58$ ($n{=}15$). Leave-one-out keeps $\rho{\in}[-0.01, +0.24]$, and dropping the two extremes (R1-Distill-32B and GPT-5) gives $\rho{=}{+}0.04$. We therefore treat coherence as a separate diagnostic, not a proxy for accuracy. The practical reading is concrete: the second- and third-most-accurate models (Opus, Sonnet) still violate complement parity on 48% and 72% of pairs, while R1-Distill-32B is among the most coherent yet ranks 9th of 15 on Brier (0.267). The per-class picture is heterogeneous (complement $\rho{=}{+}0.62$, conjunction $\rho{=}{-}0.33$, $n{=}15$), so a single composite washes out and we report all four classes (Section C).

Robustness of the closed-model estimates. A full-capture structured-output re-run of the six closed models confirms the closed-model estimates: GPT-5 violates complement parity on $10/68$ pairs, within one pair of the original count, so its low bar in Figure 1 reflects the model and not the parser (Section D).

Deployment vs. capability. A joint prompt with a sum-to-one instruction collapses the gaps to near zero (Section E). Models can be coherent when asked jointly; the violations come from the deployed interface, which elicits forecasts one event at a time, while coherence is a joint constraint. The audit thus measures deployed behavior, and the repair of Theorem 1 applies without constraint-aware prompting.

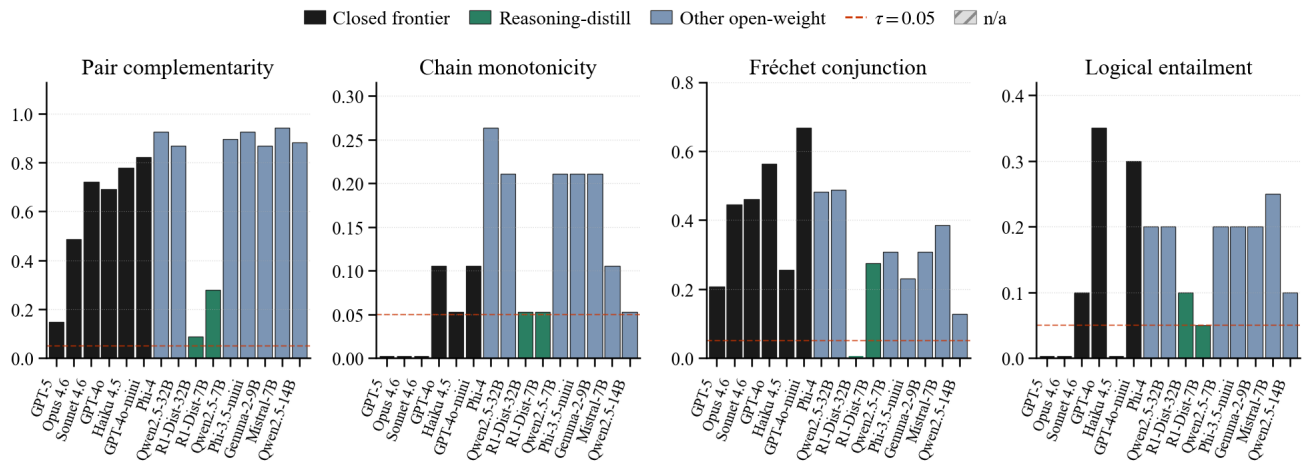

Figure 2. Per-model incoherence rates on the four classes, sorted by Brier. Flat tick at $y{=}0$: tested with 0 observed violations (Wilson 95% upper bound $\leq$18% at $n{\leq}29$); gray hatch: no coverage. Red dashed line: $\tau{=}0.05$. Per-class y-ranges differ.

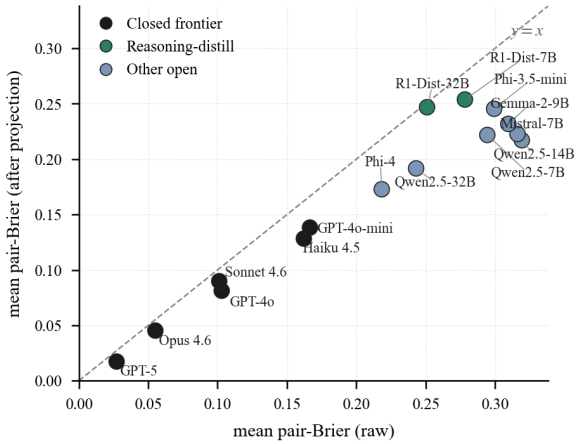

Figure 3. The coherent projection: raw vs. projected mean pair-Brier ($n{=}1005$ pairs). Aggregate recovery **20.37**%; per-pair improvement equals $g^2/4$ (Theorem 1, residual 0). Points below the $y{=}x$ diagonal indicate recovered loss.

4. Discussion

The audit complements accuracy: a model that outputs $f{\equiv}0.5$ everywhere is coherent but uninformative. The check applies wherever systems consume model probabilities before resolution, from risk monitors to market-making and prediction-market pipelines, where an inconsistent quote is a defect regardless of eventual accuracy: fed downstream, prices that fail to sum to one give conflicting expected values that no accuracy check can catch before resolution. The findings point to a simple workflow: since accuracy does not predict coherence (Section 3), coherence must be audited directly rather than read off a leaderboard, and where it fails the projection of Theorem 1 repairs complementary pairs with no extra data or model calls. Because the indicators never consult outcomes (Prop. 1), the audit is unaffected by pre-training leakage. Two checks confirm the direction (Section E): complement-violation rates are higher post-cutoff (83.1% vs 67.7%, $p{=}0.005$), and entity-masked variants show smaller gaps for 3/4 models, both opposite to a contamination story.

Repairing all four axes. Theorem 1 reprices one pair; its $g^2/4$ recovery is specific to complement parity, but every constraint class is polyhedral, so the full repair is the Euclidean projection onto their intersection by cyclic projection (Dykstra's algorithm).

Prediction-market venues. The dollar figures measure a model's internal inconsistency, not a tradeable return net of fees and liquidity. On venues such as Kalshi or Polymarket, if a model is the quoting source, complementary yes/no contracts make the gap a literal \$0–\$1 book; on forecasting platforms such as Metaculus the same number is diagnostic rather than tradeable. The coherent projection applies as a wrapper before forecasts are posted.

Limitations. $n{=}15$ forecasters bounds the precision of the accuracy-coherence association (hence a wide CI), and Opus 4.6 has partial conjunction and entailment coverage.

5. Conclusion

de Finetti's coherence is a no-arbitrage condition, so a forecaster's probabilities can be audited for incoherence without outcomes. Across 15 forecasters the implied arbitrage is large (\$0.143 per pair; 68.8% above tolerance), and a single closed-form projection recovers 20.37% of squared-error loss, exactly $g^2/4$ per pair by the complement projection identity (Theorem 1). Accuracy does not imply coherence: most forecasters, including accurate ones, quote arbitrage on a majority of pairs. We release CoherenceBench v0 and a one-command audit that flags such prices before any outcome is known.

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.

de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1): 1–68, 1937.

Delbaen, F. and Schachermayer, W. *The Mathematics of Arbitrage*. Springer Finance. Springer, 2006.

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.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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

Harrison, J. M. and Kreps, D. M. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20(3):381–408, 1979.

Harrison, J. M. and Pliska, S. R. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, 11(3):215–260, 1981.

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.

Nau, R. F. De finetti was right: Probability does not exist. *Theory and Decision*, 51(2–4):89–124, 2001.

Predd, J. B., Seiringer, R., Lieb, E. H., Osherson, D. N., Poor, H. V., and Kulkarni, S. R. Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, 55(10):4786–4792, 2009.

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

Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

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. *Science Advances*, 2024.

Seidenfeld, T., Schervish, M. J., and Kadane, J. B. The fundamental theorems of prevision and asset pricing. *International Journal of Approximate Reasoning*, 51(1): 148–158, 2010.

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*. 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. Proof of Theorem 1 and chain reprice

Proof of Theorem 1. Write $f(A)=a$, $f(B)=b$, $y(A)=\alpha$, $y(B)=\beta=1-\alpha$, $g=a+b-1$, so $P'(A)=a-g/2$, $P'(B)=b-g/2$. The per-leg delta is $(p-y)^2$ –

$(p-g/2-y)^2 = g(p-y) - g^2/4$. Summing over A, B and halving,

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

using $a+b-\alpha-\beta=(a+b)-1=g$ (as $\alpha+\beta=1$). Since $\mathrm{arb}=\frac{1}{2}|g|$, this equals arb^2. □

Chain reprice. For a chain ordered by tightness, the isotonic regression of the forecasts under the non-increasing constraint is the unique ℓ_2-projection onto the monotone cone (pool-adjacent-violators); 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 respect the chain ordering.

Proposition 1 (Outcome-free screen)**.** *Each indicator $V_{pair}, V_{chain}, V_{conj}, V_{ent}$ is a deterministic function of the forecaster's outputs and the structural specification; ground-truth outcomes appear in none. Immediate from the definitions.* □

B. Per-model arbitrage, accuracy, and loss recovery

Table 2 reports, per model (sorted by Brier): mean implied arbitrage per \$1 (1), the fraction of pairs above the τ tolerance (its complement-parity rate on the pairs it priced), the aggregate $\sum \frac{1}{2}|g|$ over its priced pairs, and the % of pair-Brier recovered by the projection of Theorem 1. Brier is pooled over events with verified outcomes by the released figure pipeline.

Table 2. Per-model static-arbitrage and projection metrics, sorted by Brier. Closed-model figures use the corrected full-capture re-run (Section D).

Model	Brier	arb/\$1	% arb	recov.
gpt-5	0.017	0.027	13%	36.5%
claude-opus-4-6	0.083	0.056	48%	17.9%
claude-sonnet-4-6	0.108	0.073	72%	11.2%
gpt-4o	0.121	0.098	70%	21.1%
claude-haiku-4-5	0.159	0.128	78%	20.9%
gpt-4o-mini	0.198	0.125	82%	17.0%
phi-4	0.222	0.178	93%	20.8%
qwen2.5-32b-awq	0.265	0.181	87%	21.2%
r1-distill-qwen-32b-awq	0.267	0.019	9%	1.7%
r1-distill-qwen-7b	0.283	0.076	28%	8.8%
qwen2.5-7b	0.291	0.228	90%	24.7%
phi-3.5-mini	0.307	0.191	93%	18.0%
gemma-2-9b	0.328	0.230	88%	25.1%
mistral-7b-v0.3	0.336	0.257	94%	29.6%
qwen2.5-14b-awq	0.355	0.276	88%	32.0%

C. Accuracy-vs-arbitrage correlation: method

Composite incoherence is the mean of a model's available class rates (per-class coverage varies; the n used is reported per class). Against mean Brier across the n=15 models we compute Spearman ρ and Pearson r; a bootstrap 95% CI resampling models with replacement (10,000 draws); a two-sided permutation p-value (10,000 permutations of the Brier vector against fixed coherence, H_0: independence); and leave-one-out. Results: ρ=+0.15, r=+0.30, CI $[-0.42, +0.65]$, p=0.58, LOO range $[-0.01, +0.24]$, drop-both ρ=+0.04. Per-class Spearman vs Brier: complement +0.62, strike +0.50, conjunction −0.33, dominance +0.34. The composite is statistically indistinguishable from zero; we report the measured statistic and decline to assert independence.

D. GPT-5 capture and the full-capture re-run

GPT-5 is a reasoning model whose chain-of-thought can exhaust the completion budget before the JSON answer. Under the original strict-JSON protocol it captured only ∼15% of GPT-5's calls (versus ∼50% for the non-reasoning OpenAI models), and on the captured calls the observed pair rate swung with parser choice (from 0% strict to 27% permissive), raising a selection-bias concern.

Re-run. We re-solicited all six closed models under a corrected protocol: provider-native structured output (OpenAI `json_schema`; Anthropic forced tool-use of a {`"probability": number`} schema) and, for the reasoning model, default reasoning effort with a 25k `max_completion_tokens` budget (a small `max_tokens` cap truncates the answer). Capture reached ∼100% (GPT-5 2619/2620 calls). GPT-5's pair rate is then 15% (10/68), within a point of the original strict-JSON 16% (11/68), so the low-capture estimate was representative rather than biased. A reasoning-effort ablation confirms coherence is reasoning-dependent: at minimal effort GPT-5's pair rate is 62%. The non-reasoning OpenAI models move negligibly under the re-run, and the Anthropic models reach 100% capture via forced tool-use.

E. Protocol and robustness

Sweep. Nine 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}) and six closed (`gpt-5`, `gpt-4o`, `gpt-4o-mini`, `claude-opus-4-6`, `claude-sonnet-4-6`, `claude-haiku-4-5`). Each $f(e)$ is the mean over N=10 samples at T=1.0 with exponential backoff (base 1.5 s, ≤ 6 retries). The system prompt is *"You are a careful probabilistic forecaster. Output a single number in [0,1], the probability the event resolves YES, with no other text."* and the user message is `Question: {q}; Resolution date: {date}; Context: {context}`, where context is the pre-resolution news

snippet for the 176 events with `corpus_available` and empty otherwise. Exact model IDs and snapshots ship in `model_versions.json`.

Deployment vs. capability. On 20 complementary pairs and four models, a joint prompt with an explicit sum-to-one instruction collapses the mean complement gap to near zero for the closed models and Qwen-2.5-7B (Table 3); deployment-realistic per-event prompting is what admits the arbitrage, and the projection of Theorem 1 recovers the same loss without constraint-aware prompting at inference time.

Table 3. Mean complement gap $|f(A)+f(B)-1|$ under separate (per-event) and joint sum-to-one prompting, on a 20-pair check.

Model	Separate	Joint+sum1
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

Leakage checks. (i) Temporal split by each model's documented training cutoff: complement-violation rates are higher post-cutoff (83.1% vs 67.7%, $p{=}0.005$) and conjunction rates lower post-cutoff, neither of which is consistent with memorized-resolution contamination inflating apparent coherence. (ii) Entity-masking on 7 pairs: $3/4$ models show smaller masked gaps (Table 4), again opposite to a recall-driven story.

Table 4. Mean complement gap on 7 entity-masked vs. original pairs. Masking shrinks the gap for $3/4$ models, the opposite of a memorized-entity story.

Model	Original	Masked
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

F. CoherenceBench v0: construction and release

The benchmark's 262 events span five domains (economics 39, geopolitics 57, politics 25, science 57, sports 84). Each event ships in `events.jsonl` with fields `event_id`, `question`, `question_open_date`, `resolution_date`, `outcome`, `domain`, `source_url`, `keywords`, the outcome verified against a cited Wikipedia revision.

Pairs (68). Disjoint outcomes sharing a resolution date (elections, finals, awards, court rulings, macro thresholds).

Chains (19, over 76 events). Tightening thresholds where the tighter event implies the looser, e.g. Bitcoin price thresholds \$2k/\$5k/\$10k/\$20k.

Conjunctions (39). Triples $(A, B, A{\cap}B)$ covering mutually exclusive cases, nested thresholds, and substantive joint events.

Entailments (21). Pairs $A{\Rightarrow}B$ where A implies B, e.g. BTC>\$20k entails BTC>\$10k.

Release. The project repository ships `events.jsonl`, the four structure-relation files, forecast logs, a sha256 manifest, and `scripts/audit.py` (CC-BY-4.0/MIT), which reproduces the scores and repair from a forecast CSV.