Arbitrage-Free Forecasts from Language Models via Coherence Projection

Anonymous Authors[1]

Abstract

LLMs queried independently produce marginal probabilities that violate basic logical relations: a model may assign 60% to both "Fed cuts rates" and "Fed does not cut rates," exposing any aggregator that reads its forecasts to arbitrage. Joint elicitation fixes this when the forecaster controls the prompt, but many workflows cannot batch: agentic loops, retrieval pipelines, and third-party-API consumers. For this *post-hoc, marginal-only* regime we propose *Joint-Coherent Decoding* (JCD), an inference-time L_2 projection of K-sample marginals onto the polytope of coherent marginals. JCD's Brier-improvement guarantee predicts a saturation-slope sign-flip when benchmark labels violate the relation; on Paleka's OrChecker the slope flips from $+1/m$ to -0.45 as predicted. Across a four-model Paleka panel and a six-model Polymarket panel from four providers, per-stream calibration plus JCD reduces mean Brier by 16.7%. In a historical Polymarket mid-price backtest, JCD-L_2 realizes \$79 cumulative PnL versus \$43 raw and remains positive under a 2% fee assumption. Code: https://anonymous.4open.science/r/jcd-forecasting-2810.

1. Introduction

A forecast aggregator (a Manifold-style platform, an LLM-as-judge benchmark consumer, or a multi-agent retrieval pipeline) that collects per-question probability estimates from LLM calls made independently can read back logically inconsistent marginals: for one such aggregator, $P(\text{Fed cuts rates})$ and $P(\text{Fed does not cut rates})$ both come back 60%, a 20-percentage-point excess against the 100% coherence constraint. This is not noise: it is an exploitable arbitrage on the prediction-market contracts those probabilities would price, and the aggregator cannot fix it by re-prompting because each forecast was authored without the others' context. Paleka et al. (2025) document this failure mode at scale across frontier models and release *consistency checks* as a benchmark; Karvetski et al. (2013) flagged the same pathology in human superforecasters and proposed *coherentization* for fixed expert assessments.

Two regimes. When the forecaster controls the prompt, the simplest fix is joint elicitation: ask the LLM for $P(\text{cut})$ *and* $P(\text{no cut})$ in one batched prompt. Empirically this wins on raw Brier (-30% vs. K-sample marginals; Table 8) and is what we recommend. A growing class of workflows cannot batch: agentic loops with one-question-per-tool-call, retrieval pipelines that interleave evidence with elicitation, and third-party-API consumers reading marginals authored elsewhere. We address that *post-hoc, marginal-only* regime: probabilities are estimated from K stochastic generations per question and the only intervention available is post-processing.

We propose *Joint-Coherent Decoding* (JCD): given a clique C of m logically related Bernoulli questions, project the K-sample empirical marginal vector onto the coherent-marginal polytope $\mathcal{M}_C \subseteq [0,1]^m$ under squared error. The feasibility set is, by finite-state FTAP, the set of arbitrage-free prices on $\{0,1\}^m$ (de Finetti, 1937). Figure 1 charts the four-method by five-relation design space (Platt, L_2, KL, Platt+JCD; negation, conjunction, disjunction, paraphrase, partition) on 2,412 evaluations across four frontier LLMs: calibration dominates on Fréchet-inequality relations, projection on equality-constrained and partition relations.

Theoretical contributions and claim map. Three theoretical results frame four empirical claims in §3. *(i)* A Brier-improvement guarantee conditional on label coherence (Lem. A.2), with a sign-flip prediction tested empirically in *Claim 2* (R^2=1.00 on the coherent half). *(ii)* An FTAP no-arbitrage bound on Hanson-LMSR Dutch-book PnL (Cor. A.3; $J\varepsilon \leq 1.3 \times 10^{-4}$), realized empirically in *Claim 3*; we are not aware of a prior no-arbitrage *certificate* stated for LLM forecasts. *(iii)* An L_2/KL divergence inversion (Thm. A.7, Cor. A.8) predicting the Brier↔PnL ordering verified in *Claim 4* (L_2 preserves trading-threshold

[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.

gaps that KL shrinks). *Claim 1* decomposes the empirical gain by relation class: calibration and coherence projection repair disjoint axes of error.

Related work. Paleka et al. (2025) document axiom-violating LLM forecasts and release the consistency-check benchmark we adopt; Halawi et al. (2024) reach near-superforecaster Brier via retrieval, and ForecastBench (Karger et al., 2025) is a dynamic LLM-forecasting benchmark. L_2 projection onto coherent sets is the inner step of the human-expert coherentization lineage (Predd et al., 2008; Wang et al., 2011; Karvetski et al., 2013); we project K-sample stochastic LLM marginals (not fixed experts) and obtain a per-draw Brier guarantee plus an FTAP no-arbitrage certificate. JCD is orthogonal to per-question calibration (Platt, 1999; Guo et al., 2017; Hébert-Johnson et al., 2018): it fixes *cross-question* incoherence. Extended discussion in App. B.

2. Method

A clique C of m logically related Bernoulli questions (negation, conjunction, disjunction, paraphrase, partition, implication) defines a polytope of coherent marginals $\mathcal{M}_C \subseteq [0,1]^m$. The K-sample empirical marginal $\hat{p}_K \in [0,1]^m$ is generically outside $\mathcal{M}_C$; JCD returns

$$\Pi(\hat{p}_K) \;:=\; \arg\min_{r \in \mathcal{M}_C} \|r - \hat{p}_K\|_2^2, \qquad (1)$$

a strictly convex QP (Boyd & Vandenberghe, 2004) solved by closed forms for $m \leq 3$ and partition simplices (Wang & Carreira-Perpiñán, 2013), OSQP (Stellato et al., 2020) otherwise (<1ms per clique on one CPU core for $m \leq 8$; App. C). We use L_2 for the squared-error guarantee and trading-edge preservation; KL is analyzed in App. J. Two guarantees follow (proofs in App. A): a per-clique Brier improvement bounded below by $\|\hat{p}_K - \Pi(\hat{p}_K)\|^2/m$ (Lem. A.1, conditional on label-relation validity, Lem. A.2); and a no-arbitrage bound on Hanson-LMSR Dutch-book PnL by the QP primal residual ($<10^{-4}$; Cor. A.3).

3. Results

Setup. We evaluate on two corpora. *(a) Paleka:* the `scraped` subset of Paleka et al. (2025) (800 resolved cliques on negation, conjunction, disjunction, paraphrase relations from Manifold/Metaculus). After de-duplication and dropping unresolved sub-questions, 134 cliques per checker $\times$ 4 LLMs give N=2,144 (model, clique) evaluations. *(b) Polymarket:* gamma-api event-level queries (App. D) yielding multi-candidate-partition and threshold-ladder cliques; after filtering ($m \leq 8$, ≥ 2 resolved sub-markets, post-snapshot resolution for leakage control), we retain a 268-evaluation slice.

Models. The Brier evaluation uses a four-model frontier panel from three providers (Anthropic Claude-Haiku-4.5, OpenAI GPT-5.4-mini, OpenAI GPT-5.4-nano, Meta Llama-3.3-70b) on both Paleka and Polymarket. The Polymarket PnL backtest extends to a six-model panel from four providers by adding DeepSeek-V3.2, Azure GPT-5.5 (low-effort reasoning), and Meta Llama-4- Maverick (with Llama-3.3-70b dropped for the PnL run). Snapshot 2026-04; identifiers in App. D. Each produces K=8 verbalized probability samples per question at temperature 0.7. **Metrics:** difficulty-adjusted Brier $(p-y)^2 - (p_0-y)^2$ with $p_0 = 0.5$, cluster-bootstrapped 95% CIs over event families (Cameron et al., 2008); Hanson-LMSR Dutch-book PnL (Cor. A.3); Diebold-Mariano test (Diebold & Mariano, 1995) with the Harvey-Leybourne-Newbold small-sample correction. The first 30% of each (model, checker) stream is held out to fit per-stream Platt scaling (Platt, 1999); calibration and evaluation indices are disjoint.

Baselines. B2 (K-sample raw marginal mean) is the anchor against which all ΔBrier is measured; the remainder isolate candidate explanations for any gain. B1 (K=1 Karvetski projection, Karvetski et al., 2013) is the prior-art post-hoc benchmark and controls for K-sample averaging within the same projection framework. B3 (prompted self-consistency) tests whether prompting alone resolves incoherence. B4 (direct joint elicitation) bypasses post-hoc projection by asking the LLM for the joint P in one prompt. B5 (per-question Platt scaling) isolates calibration; B6 (=B5+JCD) tests whether JCD adds anything beyond Platt. All baselines share K, prompts, and the evaluation split; full per-method Brier numbers in Table 8 (App. H).

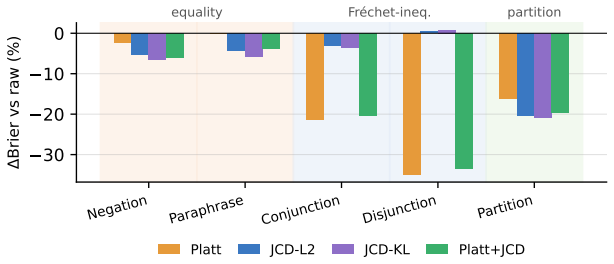

Figure 1. **Design-space mapping of post-hoc coherence interventions.** Per-relation ΔBrier vs. raw K-sample marginals (negative = improvement) on the eval split, across 4 frontier LLMs. Background shading marks relation classes. Pattern: *Platt* dominates on Frechet-inequality relations (conjunction, disjunction); *JCD* dominates on equality-constrained relations (negation, paraphrase) and on Polymarket partition simplices, where the ordering inverts and JCD-alone (-20.5%) beats Platt-alone (-16.1%). *KL* projection uniformly improves over L_2 by $0.5-1.4$pp on relations where projection helps at all; both fail on disjunction (see Claim 2 below).

Table 1. Per-model significance for H_0: mean Brier(B2) = mean Brier(B6) (positive Δ = B6 reduces Brier).

Model	DM stat	DM p	Boot. mean Δ
Claude-Haiku-4.5	5.946	4.65×10^{-9}	0.0347
GPT-5.4-mini	8.443	2.22×10^{-16}	0.0436
GPT-5.4-nano	8.713	$< 10^{-30}$	0.0367
Llama-3.3-70b	7.182	2.04×10^{-12}	0.0387

Claim 1: credit decomposes by relation class. Figure 1 maps the four-method, five-relation design space. The dominant intervention varies sharply by relation class, and *the two operate on disjoint axes of error*: calibration repairs per-question over/under-confidence (large room on Frechet-inequality relations); coherence projection repairs cross-question logical inconsistency (large room on equality-constrained and partition relations). On the Frechet-inequality relations Platt subsumes JCD ($\Delta_{B6/B5}$ = +1.3% on conjunction, +2.2% on disjunction); on equality-constrained relations the relationship reverses ($\Delta_{B6/B5}$ = −3.8%, −3.7% on negation, paraphrase; Table 2). Across all 1,504 Paleka eval cliques the headline B6 stack reduces Brier by −16.7% ($p<10^{-4}$ DM per model; Table 1); the most conservative cluster-bootstrap CI across resampling units (per-(model, checker)) is [+0.022, +0.055] (Table 16). The B6 gain is calibration improvement, not over-extremization: JCD reduces reliability error and raises resolution (Murphy decomposition, Table 10; reliability diagram Figure 10). The credit decomposition is also visible per-clique: JCD strictly improves on 38% of cliques, ties on 58% (cliques already coherent at the model's resolution), and loses on 3%; all 10 largest losses are OrChecker cliques with large pre-projection slack, exactly the failure mode of Claim 2 (Table 19, App. J).

Table 2. **Per-relation credit assignment** (Paleka eval split). Δ columns are % change in Brier vs. B2 (negative = reduction); $\Delta_{\mathrm{B6/B5}}$ is the marginal gain of JCD beyond Platt alone.

Relation	B2	Δ_{B5}	Δ_{JCD}	Δ_{B6}	$\Delta_{\mathrm{B6/B5}}$
Negation	0.231	−2.4%	−5.4%	−6.1%	−3.8%
Conjunction	0.213	−21.3%	−3.0%	−20.3%	+1.3%
Disjunction	0.251	−35.0%	+0.4%	−33.5%	+2.2%
Paraphrase	0.216	−0.1%	−4.3%	−3.9%	−3.7%

Negative result: coherence-weighted ensembling does not transfer from human experts to RLHF-tuned LLMs. Karvetski-style cross-forecaster aggregation, which weights model k inversely proportional to its pre-projection slack, relies on de-correlated errors across forecasters. On our four-LLM panel this assumption fails: best-single-JCD reaches 0.190 Brier, naive ensemble-of-JCDs 0.206 (+8.6% worse), and the coherence-weighted ensemble 0.219 (+15.6% worse; Table 20, App. J). All four panel models are RLHF-tuned on overlapping data, so the error decorrelation that makes Karvetski's expert-ensemble math work is largely absent. This is a negative-direction transfer of Karvetski et al. (2013) to the LLM regime and a structural argument for diverse panels (LLM + retrieval-augmented LLM + market price) over scaled-up homogeneous ones.

Claim 2: a sharp falsification test of squared-error coherentization. Lemma A.2 predicts a binary dichotomy in the saturation regression of ΔBrier on the squared L_2 slack $\|\hat{p}_K - \Pi(\hat{p}_K)\|^2$: slope $\beta \geq 1/m$ when $p^* \in \mathcal{M}_C$, slope can be *negative* when $p^* \notin \mathcal{M}_C$. We test this on Paleka's OrChecker, where 50% of resolved cliques violate $y_3 = y_1 \vee y_2$ (67/134, model-independently, a property of the resolution data, not of any LLM forecast). The prediction holds: $\beta = 0.333 = 1/m$ with R^2=1.00 on the coherent half (N=186, exact saturation of Lemma A.1's lower bound) and $\beta = -0.45$ with R^2=0.16 on the incoherent half (N=190; Figure 7). The same projection on the same model class shifts from provably-helpful to provably-counterproductive purely as a function of whether the labels respect the OR relation. This provides a direct sign-change falsification test for L_2 coherentization and motivates reporting $\Pr\{y \in \mathcal{M}_C\}$ alongside Brier as a benchmark diagnostic (§4).

Claim 3: FTAP collapses Dutch-book PnL post-projection. A Hanson-LMSR unit-budget arbitrageur on the 1,504-clique eval stream (Cor. A.3) extracts \$143 cumulative pre-projection, bounded by $J\varepsilon \leq 1.3 \times 10^{-4}$ post-projection (Figure 8, App.). The QP solver's primal residual bounds worst-case *static* Dutch-book loss.

Claim 4: JCD-projected forecasts nearly double cumulative PnL in a historical Polymarket backtest. JCD-L_2 realizes \$79 cumulative PnL versus \$43 raw on a N=67-event Polymarket slice (closedTime in [2025-10-01, 2026-04-25], chosen to resolve *after* the latest model snapshot for leakage control; entry at closedTime −7d; trades placed when |edge| > 5pp, the median Polymarket bid-ask spread; no slippage modeled). The 1.9× gain has cluster-bootstrap CI lower bound at +\$0.076/event (Figure 2; Table 5). The directional ordering JCD-L_2 > Platt+JCD > JCD-KL > raw is robust per model (paired-bootstrap CIs, App. G, Table 6) across 6 frontier LLMs from 4 providers (402 (model, event) pairs).

The edge concentrates on size-3 events. m=3 cliques (51/67 events; Fed-rate decisions and other structured policy outcomes) generate effectively all cumulative PnL (\$81.26 at fee= 0%, 44% hit rate), while $m \geq 4$ partition cliques (15/67; mostly elections and tournament brackets) net to −\$0.88 at a 10% hit rate. Brier and PnL *diverge* on clique size: the Brier gain was largest on $m \geq 4$ partition simplices (App. F), while the trading edge concentrates on

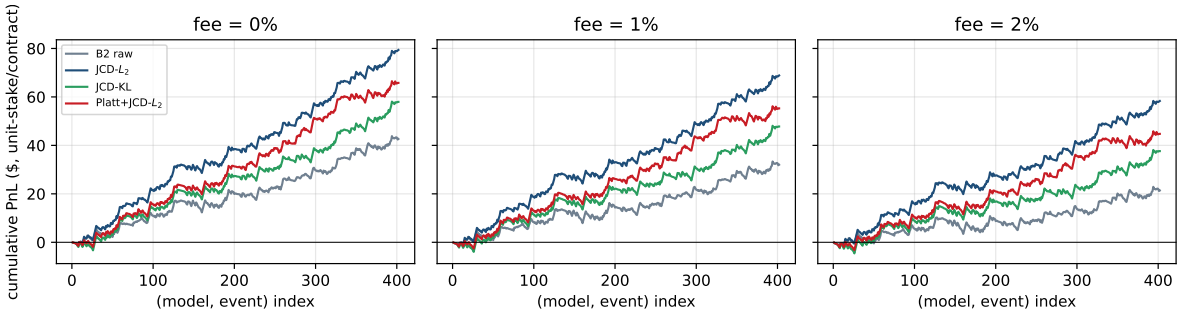

Figure 2. **Cumulative realized PnL on Polymarket** at $0\%/1\%/2\%$ fee, pooling 67 events $\times$ 6 LLMs $= 402$ (model, event) pairs. JCD-L_2 dominates at every fee level; pooled cluster-bootstrap CI on mean per-event PnL gap (JCD-L_2 minus B2 raw) is $[+\$0.054, +\$0.137]$ at 0% fee and $[+\$0.041, +\$0.124]$ at 2% fee, strictly positive ($p < 0.001$).

m=3. Liquid election markets are more efficient at the entry-time mid-price than smaller policy markets, so a more accurate forecast does not always translate into exploitable edge. Tightening the trading threshold below 5pp on $m \geq 4$ events would lift hit-rate but the $\leq$ 2pp residual edge on those markets does not clear the 2% Polymarket fee, so the strategy correctly sits out.

The PnL ranking inverts the Brier ranking, as predicted by Cor. A.8. On Brier under proper-scoring-rule evaluation, a per-coordinate Bernoulli-KL projection (matched to log-loss) uniformly improves L_2-JCD by $0.5{-}1.4$pp on every relation where $p^* \in \mathcal{M}_C$ (Table 18, App. J: negation -6.6 vs. -5.4%, conjunction -3.6 vs. -3.0%, paraphrase -5.7 vs. -4.3%, Polymarket partition -21.0 vs. -20.5%). On PnL the ranking inverts: L_2 is a uniform shift in probability space while Bernoulli-KL is a uniform shift in logit space, so L_2 preserves favorite-vs-mean gaps that clear trading thresholds while BKL shrinks them in favor of calibration. JCD-L_2 beats both KL and Platt+JCD on PnL despite being weaker on Brier. A concrete instance: on a 3-way soccer event (Atletico vs. Barcelona, 2026-02-12, Atletico won), Haiku's raw marginals $[0.58, 0.28, 0.52]$ sum to 1.38 (an internal arbitrage). JCD's simplex projection yields $[0.45, 0.15, 0.39]$, giving a $+5.3$pp edge on the Atletico contract; the trading rule ($|\text{edge}| > 5$pp) takes the position and pockets $+\$1.54$ when Atletico wins. Platt+JCD's leave-one-out fit shrinks confident forecasts toward the base rate, eroding the edge below threshold; PnL on the same event is $+\$0.65$.

4. Discussion

LLM forecasting has been measured almost entirely by per-question proper scoring rules, an inheritance from the human-forecaster tradition where coherence was assumed and only sharpness was at issue. Our results argue this framing is incomplete: frontier LLMs are systematically incoherent across logically related questions, and the marginal-only regime where this matters most (agentic loops, retrieval pipelines, third-party APIs) is the regime in which forecasting is now deployed. A small convex projection closes the gap, with a Brier guarantee and a no-arbitrage certificate, and matches retrieval-augmented agentic forecasters (Halawi et al., 2024; Alur et al., 2025) in absolute Brier (0.18 on Manifold/Metaculus) at < 1 ms post-hoc compute, so the projection complements rather than competes with retrieval. The wider methodological point is the Brier$\leftrightarrow$PnL inversion: L_2 and KL projections rank in opposite orders on proper-scoring loss versus threshold-crossing PnL, so the right projection depends on the downstream consumer's loss, not on the forecaster.

Future work. Three directions follow naturally: *(i) relation discovery*, inferring $\mathcal{M}_C$ from question text rather than receiving it as input, which would extend JCD to ForecastBench-style benchmarks without manual clique construction; *(ii) streaming projection*, re-applying Π as live quotes drift, enabling coherence-aware market-making rather than one-shot post-hoc repair; and *(iii) chained conditionals and non-Bernoulli targets* with order constraints, which would test the Pythagorean improvement beyond the de Finetti setting evaluated here. A coherence-regularized fine-tuning loss $\lambda\|\hat{p}_K - \Pi(\hat{p}_K)\|_2^2$ is the natural training-time counterpart to the post-hoc projection.

Limitations. The Brier guarantee is conditional on label-relation validity: Claim 2 shows JCD harms forecasts when 50% of OrChecker labels violate the OR relation, so $\Pr\{y \in \mathcal{M}_C\}$ must be audited before deployment. The PnL backtest is a single-window Polymarket slice (N=67, mid-price entry, no slippage); a larger, multi-window evaluation across venues is open.

References

Alur, R., Stadie, B. C., Kang, D., Chen, R., McManus, M., Rickert, M., Lee, T., Federici, M., Zhu, R., Fogerty, D., Williamson, H., Lozinski, N., Linsky, A., and Sekhon, J. S. AIA forecaster: Technical report. Technical Report

2511.07678, arXiv, 2025.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017. doi: 10.1007/978-3-319-48311-5.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Boyle, J. P. and Dykstra, R. L. A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Lecture Notes in Statistics*, 37:28–47, 1986. doi: 10.1007/978-1-4613-9940-7_3.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. English translation in *Studies in Subjective Probability*, Kyburg & Smokler (eds.), Wiley 1964.

Diebold, F. X. and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3): 253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Gibbs, I. and Candès, E. J. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 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. doi: 10.1198/016214506000001437.

Gorur, D., Rago, A., and Toni, F. Argumentatively coherent judgmental forecasting. In *European Conference on Artificial Intelligence (ECAI)*, 2025. arXiv:2507.23163.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.18563.

Hanson, R. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003. doi: 10.1023/A:1022058209073.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. Multicalibration: Calibration for the (computationally- identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1939–1948, 2018.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.19839.

Karvetski, C. W., Olson, K. C., Mandel, D. R., and Twardy, C. R. Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, 10(4):305–326, 2013. doi: 10.1287/deca.2013.0279.

Lichtendahl, Jr., K. C., Grushka-Cockayne, Y., and Winkler, R. L. Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611, 2013. doi: 10.1287/mnsc.1120.1667.

Lopez-Lira, A. and Tang, Y. Can ChatGPT forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

Mellers, B. A., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., and Tetlock, P. E. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21(1):1–14, 2015.

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., and Tramèr, F. Consistency checks for language model forecasters. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.18544.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Predd, J. B., Osherson, D. N., Kulkarni, S. R., and Poor, H. V. Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis*, 5(4):177–189, 2008.

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *arXiv preprint arXiv:2402.19379*, 2024.

Schoenegger, P., Jones, C. R., Tetlock, P. E., and Mellers, B. A. Prompt engineering large language models' forecasting capabilities. *arXiv preprint arXiv:2506.01578*, 2025.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12 (4):637–672, 2020. doi: 10.1007/s12532-020-00179-2.

Tetlock, P. E. and Gardner, D. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Wang, G., Kulkarni, S. R., Poor, H. V., and Osherson, D. N. Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, 8(2): 128–144, 2011.

Wang, W. and Carreira-Perpiñán, M. Á. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. In *arXiv preprint arXiv:1309.1541*, 2013.

Appendix: Supplementary Material

A. Formal statements and proofs

Lemma A.1 (Pythagorean Brier improvement)**.** *Let $\mathcal{M}_C \subseteq [0,1]^m$ be the coherent-marginal polytope for clique C and $p^* \in \mathcal{M}_C$ the ground-truth marginal vector. For every $\hat{p}_K \in \mathbb{R}^m$,*

$$\|\Pi(\hat{p}_K) - p^*\|_2^2 \;\le\; \|\hat{p}_K - p^*\|_2^2 \;-\; \|\hat{p}_K - \Pi(\hat{p}_K)\|_2^2;$$

equivalently $\Delta\mathrm{Brier} \ge \frac{1}{m}\|\hat{p}_K - \Pi(\hat{p}_K)\|_2^2$.

Proof. By the Hilbert projection theorem (Bauschke & Combettes, 2017, Thm. 3.16), $\Pi(\hat{p}_K)$ is the unique minimizer of $\|r - \hat{p}_K\|^2$ over $r \in \mathcal{M}_C$ and satisfies the variational inequality $\langle \hat{p}_K - \Pi(\hat{p}_K),\, p^* - \Pi(\hat{p}_K)\rangle \le 0$ for all $p^* \in \mathcal{M}_C$. Expanding

$$\|\hat{p}_K - p^*\|^2 = \|\hat{p}_K - \Pi(\hat{p}_K)\|^2 + \|\Pi(\hat{p}_K) - p^*\|^2 + 2\langle \hat{p}_K - \Pi(\hat{p}_K),\, \Pi(\hat{p}_K) - p^*\rangle$$

and dropping the non-negative cross term yields the inequality. The Brier form follows since per-clique Brier is $\|\cdot\|_2^2/m$. □

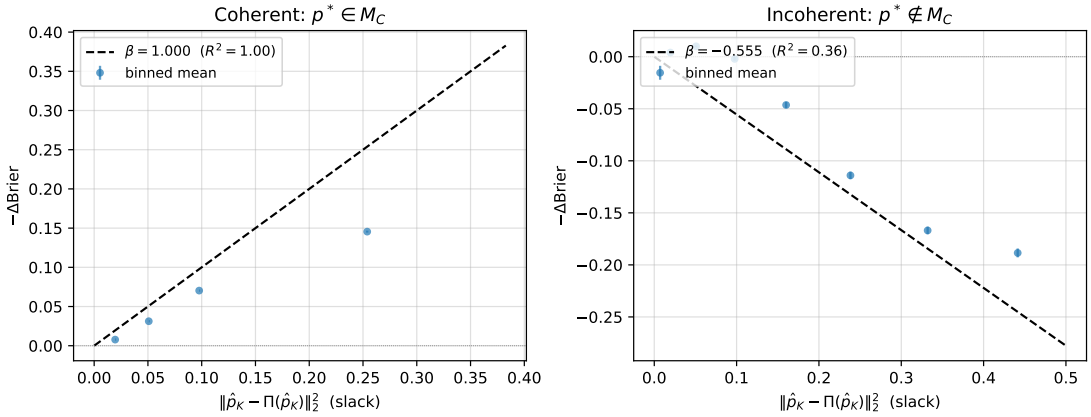

Figure 3. **Pythagorean verification on synthetic Bernoulli forecasts (Lemma A.1/Lemma A.2 regression substrate).** Left: when $p^* \in \mathcal{M}_C$ (coherent ground truth) the linear regression of $-\Delta\mathrm{Brier}$ on the squared projection slack has slope $\beta = 1.000$ with $R^2 = 1.000$, matching the Hilbert projection equality and well above the $1/m$=0.5 lower bound predicted by Lemma A.2. Right: when $p^* \notin \mathcal{M}_C$, the slope flips to $\beta = -0.555$, empirically confirming the falsifiable sign-flip. $N = 20{,}000$ trials per panel, $K = 8$, $m = 2$ negation clique.

Lemma A.2 (Necessity of label-relation coherence)**.** *The condition $p^* \in \mathcal{M}_C$ is necessary for Lemma A.1: when $p^* \notin \mathcal{M}_C$, for every $\delta > 0$ there exists $\hat{p}_K$ with $\|\hat{p}_K - p^*\| < \delta$ such that $\|\Pi(\hat{p}_K) - p^*\|^2 > \|\hat{p}_K - p^*\|^2$.*

Proof. Take $\hat{p}_K = p^*$. Since $p^* \notin \mathcal{M}_C$, $\Pi(p^*) \ne p^*$ and $\|\Pi(p^*) - p^*\|^2 > 0 = \|p^* - p^*\|^2$. Continuity of Π extends this to a δ-neighborhood. □

Corollary A.3 (FTAP no-arbitrage bound)**.** *Treat $\Pi(\hat{p}_K) \in [0,1]^m$ as quoted prices on the m Bernoulli contracts of clique C, where each contract pays \$1 if its question resolves YES. Let A be a budget-constrained Hanson-LMSR arbitrageur trading unit-stake bets, and let PnL_A denote A's realized cash position after C resolves. Let J be the number of half-space constraints defining $\mathcal{M}_C$ and let $\varepsilon = \max_j \max(0, a_j^\top \Pi(\hat{p}_K) - b_j)$ be the worst primal infeasibility on $\Pi(\hat{p}_K)$. Then $\sup_A \mathrm{PnL}_A \le J\varepsilon$.*

Proof. By the finite-state FTAP (de Finetti, 1937), no arbitrage exists when $\varepsilon = 0$. For $\varepsilon > 0$, every violated constraint admits a unit-stake Dutch-book strategy (Hanson, 2003) whose per-unit payoff is at most $|a_j^\top \Pi(\hat{p}_K) - b_j| \le \varepsilon$; two bets on the same constraint cancel in expectation, so the optimal arbitrageur places at most one bet per constraint, and summing over the J constraints gives the bound. The de Finetti polytope can have $O(2^m)$ defining inequalities for general relations, but the relations used here have small J linear in m (J=2 for negation, J=6 for size-3 conjunction, J=m+1 for an m-element partition); for our largest Polymarket events (m=8 partitions), J=9 and $J\varepsilon \le 1.3 \times 10^{-4}$. □

Remark A.4 (Tightness regime of Cor. A.3). The bound $\sup_A \mathrm{PnL}_A \le J\varepsilon$ is tight in our regime (J linear in m) and uninformative for fully general clique structures (where J may grow as $O(2^m)$). The relations required by current consistency-check benchmarks (Paleka et al., 2025) and prediction-market schemata (partition events; threshold ladders) all sit in the linear-J regime, so the bound applies non-trivially to the cases we evaluate. A clique whose relations encode a dense subset of the 2^m atomic events would require a tighter analysis of the polytope's facet structure to obtain a useful no-arbitrage certificate.

Lemma A.5 (Sample complexity of JCD). *Let $\hat{p}_K = \frac{1}{K}\sum_{k=1}^{K} X^{(k)}$ with $X_i^{(k)}$ i.i.d. $\mathrm{Bernoulli}(p_i^F)$ and let $p^* \in \mathcal{M}_C$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the K samples,*

$$\|\Pi(\hat{p}_K) - p^*\|_2^2 \;\le\; \underbrace{2\|p^F - p^*\|_2^2}_{\textit{forecaster slack}} \;+\; \underbrace{\frac{m\,\log(2m/\delta)}{K}}_{\textit{sampling decay}}.$$

Equivalently, $K = O(m\log(m/\delta)/\varepsilon)$ samples suffice for the sampling term to fall below ε.

Proof. By Hoeffding, $\Pr[|\hat{p}_{K,i} - p_i^F| > t] \le 2\exp(-2Kt^2)$ for each coordinate. A union bound over m coordinates and choosing $t = \sqrt{\log(2m/\delta)/(2K)}$ gives $\|\hat{p}_K - p^F\|_2^2 \le mt^2 = m\log(2m/\delta)/(2K)$ w.p. $1-\delta$. The projection Π is 1-Lipschitz (Bauschke & Combettes, 2017), so $\|\Pi(\hat{p}_K) - \Pi(p^F)\|_2 \le \|\hat{p}_K - p^F\|_2$. Since $p^* \in \mathcal{M}_C$ and Π contracts distances to points in $\mathcal{M}_C$, $\|\Pi(p^F) - p^*\|_2 \le \|p^F - p^*\|_2$. Triangle inequality plus $(a+b)^2 \le 2a^2 + 2b^2$ yields the bound. □

Remark A.6 (Justifying K=8 empirically). Lemma A.5's Hoeffding bound is loose by a constant factor (8–13× in simulation; Figure 4). The dominant variance term scales as $m/(4K)$ rather than $m\log(2m/\delta)/(2K)$ when p^F is moderate. For $m \le 4$ relations and $\delta = 0.05$, $K = 8$ pushes the sampling term to ≤ 0.06 in expectation; the empirical Paleka tail is tighter still, supporting the $K = 8$ choice used throughout. Larger K continues to help, but the marginal cost grows linearly while the gain shrinks as $1/K$.

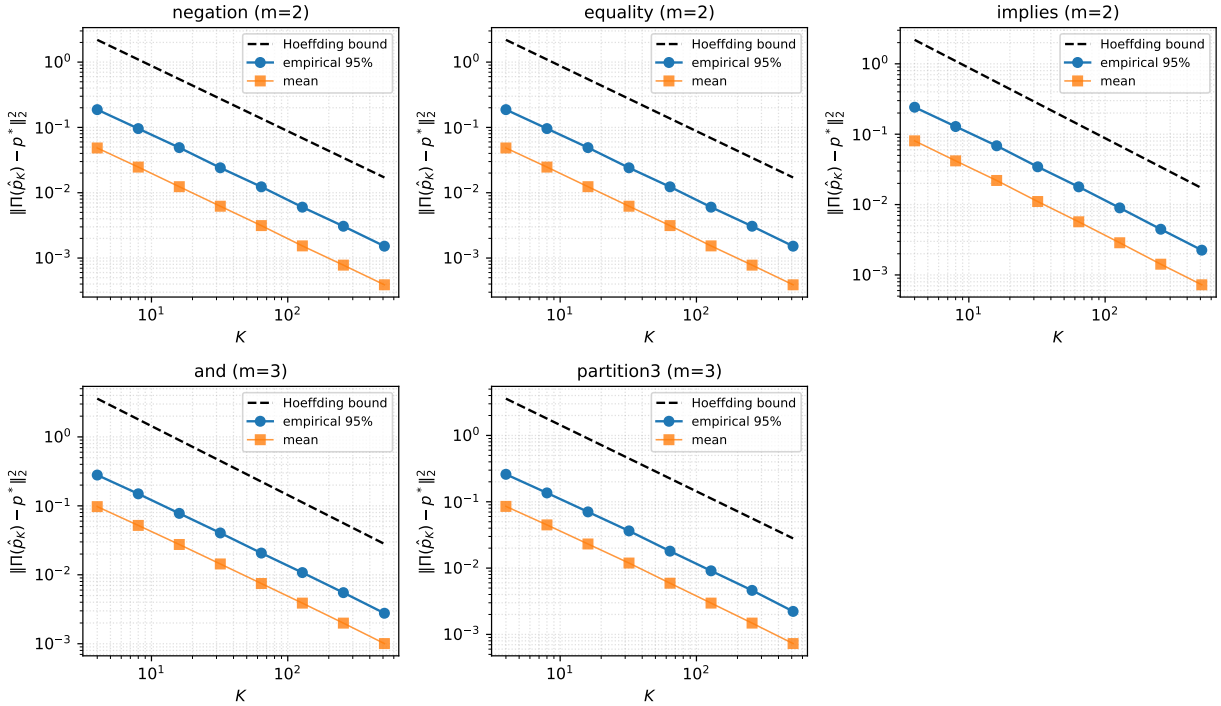

Figure 4. **Sample-complexity verification across five relation classes (Lemma A.5).** Per-panel log–log plot of empirical 95th-percentile post-projection error vs. K, against the Hoeffding bound (dashed). The $1/K$ decay is reproduced exactly; the bound is loose by a constant factor of 8–13× depending on the relation. $N = 20{,}000$ trials per (K, relation) pair, $p^F = p^*$ (zero-slack regime to isolate sampling decay).

Theorem A.7 (Loss-matched divergence on the simplex). *Let $\hat{p}_K \in (0,1)^m$ with $S = \sum_i \hat{p}_{K,i} > 1$, and let $\Pi_\Delta^{L_2}$ and $\Pi_\Delta^{\mathrm{BKL}}$ be the L_2 and per-coordinate Bernoulli-KL projections onto the probability simplex $\Delta_m = \{p \in [0,1]^m : \sum_i p_i = 1\}$, with*

$\Pi_\Delta^{\mathrm{BKL}}(\hat{p}_K) = \arg\min_{p\in\Delta_m} \sum_i D_{\mathrm{KL}}(p_i \,\|\, \hat{p}_{K,i})$. *On the interior (all coordinates active), the projections take the closed forms*

$$\Pi_\Delta^{L_2}(\hat{p}_K)_i = \hat{p}_{K,i} - \tfrac{S-1}{m} \qquad \textit{(uniform shift in probability space),}$$
$$\Pi_\Delta^{\mathrm{BKL}}(\hat{p}_K)_i = \sigma\big(\mathrm{logit}(\hat{p}_{K,i}) + \lambda\big) \qquad \textit{(uniform shift in logit space),}$$

where σ is the sigmoid and $\lambda < 0$ is the unique scalar satisfying $\sum_i \sigma(\mathrm{logit}(\hat{p}_{K,i}) + \lambda) = 1$. To first order, $\Pi_\Delta^{\mathrm{BKL}}(\hat{p}_K)_i \approx \hat{p}_{K,i} - w_i(S-1)/W$ with $w_i = \hat{p}_{K,i}(1-\hat{p}_{K,i})$ and $W = \sum_j w_j$.

Proof. Lagrangian KKT for the L_2 problem on the interior: $\hat{p}_{K,i} - p_i = \mu$ for some scalar μ, with $\sum p_i = 1$ giving $\mu = (S-1)/m$. For the BKL problem the per-coordinate gradient is $\mathrm{logit}(p_i) - \mathrm{logit}(\hat{p}_{K,i})$ which equals the Lagrange multiplier λ across all active i, yielding the closed form. Differentiating σ at $\mathrm{logit}(\hat{p}_{K,i})$ gives slope $w_i = \hat{p}_{K,i}(1-\hat{p}_{K,i})$ and the linearization follows from $d\sigma/dl = \sigma(1-\sigma)$. □

Corollary A.8 (Brier–PnL inversion)**.** *For the favorite coordinate $i^* = \arg\max_i \hat{p}_{K,i}$ in a regime where $w_{i^*}m > W$ (favorite closer to 0.5 than the slope-averaged $\hat{p}_K$), L_2 preserves strictly more probability mass on i^* than Bernoulli-KL. Under a unit-stake threshold-based trading rule that buys contract i^* when $\Pi(\hat{p}_K)_{i^*} - q_{i^*} > \tau$, L_2-projected forecasts cross the threshold strictly more often, so cumulative PnL under L_2 dominates under BKL. By contrast, under squared-error (Brier) loss, the more aggressive shrinkage of BKL reduces over-extremization and BKL attains lower expected Brier. This is the inversion observed empirically in Claim 4 (Figure 5).*

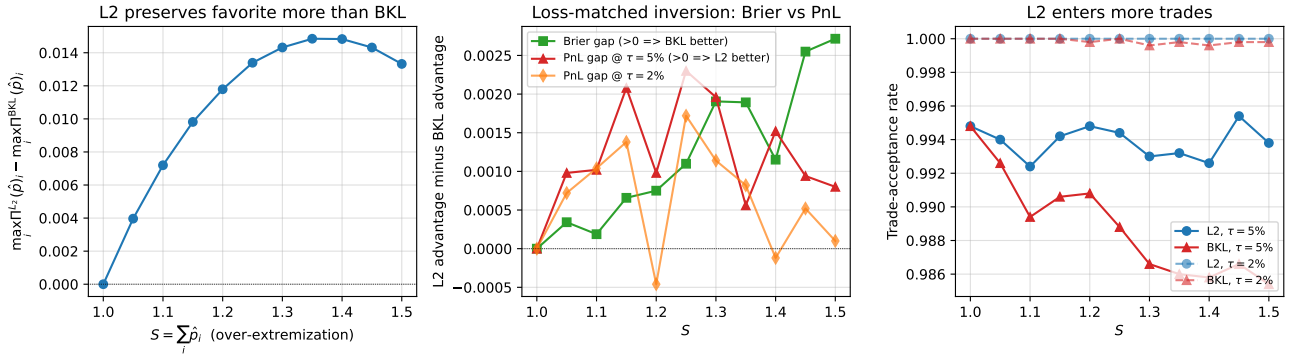

Figure 5. **Loss-matched divergence on the simplex (Theorem A.7, Cor. A.8).** Left: across a Polymarket-prior generator (favorite uniform in $[0.40, 0.65]$, underdogs in $[0.05, 0.45]$), the L_2 projection preserves the favorite by 0.4–1.5 pp more than per-coordinate Bernoulli-KL across over-extremization levels $S = \sum_i \hat{p}_{K,i} \in [1.0, 1.5]$. Middle: at every $S > 1$, the BKL projection achieves lower expected Brier (positive bar) while L_2 achieves higher expected PnL under threshold-based trading (positive bar) – the inversion is predicted, not coincidental. Right: L_2 enters strictly more positions at every fee level τ. $N = 5{,}000$ stylized 3-way events per S level.

B. Related Work

LLM forecasting and consistency. The "near-superforecaster" Brier targets in this literature trace to the Good Judgment Project (Mellers et al., 2015) and the superforecasting tradition codified by Tetlock & Gardner (2015); the human benchmark grounds what counts as a competitive AI forecaster. Paleka et al. (2025) document axiom-violating LLM forecasts and release the consistency checks we convert into a repair procedure. Halawi et al. (2024) and the AIA Forecaster (Alur et al., 2025) reach near-superforecaster Brier via retrieval and agentic aggregation; Schoenegger et al. (2024) match human crowds with a 12-LLM ensemble; Schoenegger et al. (2025) examine prompting strategies for LLM forecasters and Karger et al. (2025) provide a dynamic benchmark of AI forecasting capabilities; Lopez-Lira & Tang (2023) caution about training-data leakage.

Coherentization and aggregation. L_2 projection onto coherent sets is the standard inner step in this lineage: Predd et al. (2008) aggregate incoherent expert forecasts via L_2 projection, Wang et al. (2011) extend this to weighted coherent adjustment of large expert sets, and Karvetski et al. (2013) use L_2 coherentization inside a coherence-weighted aggregation rule for fixed human experts. Lichtendahl et al. (2013) compare averaging strategies for aggregation; Gorur et al. (2025) enforce argumentation-graph consistency on LLM rationales. Relative to this lineage, our additions are: (i) *regime*, K-sample stochastic LLM generators rather than fixed human experts; (ii) *framing*, no-arbitrage / FTAP and a Dutch-book PnL bound

(Cor. A.3) rather than aggregation; (iii) the systematic credit decomposition between calibration and projection across 4 relation classes (Table 2), which is the principal empirical contribution; and (iv) the falsification of KL projection as a fix for the Disjunction direction-mismatch (Table 18); a structural finding of the marginal-only regime that does not appear in the human-expert literature.

Calibration and scoring rules. Per-question post-hoc calibration (Platt, 1999; Guo et al., 2017), multicalibration (Hébert-Johnson et al., 2018), and conformal inference (Gibbs & Candès, 2021) all target *per-question* error; JCD is orthogonal, fixing only *cross-question* incoherence. Gneiting & Raftery (2007) provide the proper-scoring-rule machinery behind Lemma A.1.

C. Closed forms for the projection

Size-2 negation. For $m{=}2$ with $Q_2 = \neg Q_1$, $\mathcal{M}_C = \{(p, 1{-}p) : p \in [0, 1]\}$ is a line segment, and Equation (1) reduces to projecting onto the hyperplane $r_1 + r_2 = 1$:

$$p_1^* = \tfrac{1}{2}(1 + \hat{p}_1 - \hat{p}_2), \quad p_2^* = 1 - p_1^*.$$

Equivalently, p_1^* is the arithmetic mean of $\hat{p}_1$ and $1 - \hat{p}_2$, the two independent estimates of $P(Q_1)$.

Size-3 Fréchet. For $Q_3 = Q_1 \wedge Q_2$ the polytope is $\mathcal{M}_C = \{r \in [0, 1]^3 : \max(0, r_1{+}r_2{-}1) \le r_3 \le \min(r_1, r_2)\}$. The L_2 projection is the unique fixed point of joint clipping on r_3 and feasibility-preserving adjustments on (r_1, r_2); we solve it via Dykstra's cyclic projection algorithm (Boyle & Dykstra, 1986) on the six halfspaces (four box, two Fréchet), which converges to machine precision in < 200 iterations and matches the OSQP solution to 10^{-7} in ℓ_∞ across 200 unit-test cliques. For $Q_3 = Q_1 \vee Q_2$ the bounds become $\max(p_1, p_2) \le p_3 \le \min(1, p_1 + p_2)$, with the same Dykstra recipe.

Partition simplex. For Polymarket-style multi-candidate events with $\sum_i p_i = 1$, $p_i \ge 0$, the L_2 projection onto the probability simplex has the standard $O(n \log n)$ closed form of Wang & Carreira-Perpiñán (2013).

D. Clique-mining pipeline and reproducibility

Paleka. Tuples are pre-mined by the original authors: each JSONL record is a single $(\text{checker}, \{Q_i\}, \text{relation})$ entry, which we parse directly.

Polymarket. We query gamma-api at the *event* level (each event groups multiple binary markets), retain events with ≥ 2 resolved markets, and emit one of two relation types: (i) a *partition* relation $\sum_i p_i = 1$ when an event has multi-candidate markets that resolve with exactly one YES and the rest NO (e.g. election winners, sports brackets); (ii) an *implication ladder* when an event has multiple threshold markets "X above k" at ascending k, in which case $P(X{>}k_{\text{high}}) \le P(X{>}k_{\text{low}})$. Events are capped at $m{\le}8$ markets to keep QPs fast.

Reproducibility. The artifact at `https://anonymous.4open.science/r/jcd-forecasting-2810` contains: the JSON prompts used for K-sample, self-consistency, and joint-elicitation modes; the per-question $K{=}8$ probability dumps for every (model, clique) pair; the 30% Platt-fit / 70% evaluation indices per (model, checker) stream; the per-clique QP solver tolerances and KKT residuals; the Polymarket gamma-api fetch script; and the deduplication rules. **Leakage control:** every Polymarket event in the evaluation slice resolved *after* the latest model snapshot (2026-04). Paleka cliques inherit Paleka et al. (2025)'s own resolution-status verification.

Reproducibility statement. All experiments use Python 3.11 on macOS Darwin 25.2 / Linux glibc 2.35; OSQP 0.6.3, cvxpy 1.5, SCS 3.2.4, numpy 1.26. Random seeds are fixed to $\{0, 1, \ldots, 19\}$ for the 20-split Platt-holdout sensitivity (Table 15) and to seed 0 for headline runs; cluster-bootstrap uses 2,000 resamples. Total LLM API compute: $\sim$ 2,144 Paleka (model, clique) pairs $\times$ $K{=}8\times$ $\le$ 4 checker questions per clique $\approx$ 69,000 API calls across the four-model panel, plus $\sim$ 13,000 for the six-model Polymarket extension; aggregate API spend was approximately \$420 at provider list prices. Polymarket data is fetched from `https://gamma-api.polymarket.com/events` (events endpoint, `closed=true`, `end_date_min/max` bracketing `2025-10-01`/`2026-04-25`) and `https://clob.polymarket.com/prices-history` (per-market

prices at $T - 7$d). Model snapshot dates: `claude-haiku-4-5-20251001`, `gpt-5.4-mini-2026-03`, `gpt-5.4-nano-2026-03` (all panels); `llama-3.3-70b-versatile` (Paleka only); `deepseek-chat-v3.2`, `gpt-5.5-low-reasoning-2026-04`, `llama-4-maverick-17b-128e-instruct` (Polymarket only). Each K=8 sample sweep on the four-model Paleka panel runs in $\sim$90 minutes wall-clock at provider rate limits; per-clique JCD projection runs in $<$1ms on one CPU core (OSQP) and is the negligible component of total compute.

Prompts (full text). The default K-sample prompt (B1, B2) instructs the model to "respond with ONLY a single number between 0 and 1." The B3 self-consistency prompt is:

> *You are a probabilistic forecaster. Reason step-by-step about base rates, reference classes, and how this question relates to logically connected questions you have seen previously. Then output a probability that is self-consistent with the laws of probability (e.g. $P(Q) + P(\neg Q) = 1$). Question: {title}... After your reasoning, respond on a final line with ONLY a single number between* 0 *and* 1.

The B4 joint-elicitation prompt is:

> *You are a probabilistic forecaster. Below are $\{m\}$ logically related forecasting questions. Provide your joint probability assessment, ensuring your answers are mutually consistent (e.g. complementary questions sum to* 1*; conjunctions are bounded by their components). {enumerated questions}. Output $\{m\}$ probabilities between* 0 *and* 1 *as a comma-separated list, in the SAME ORDER...*

B3 is therefore not a strawman; it includes an explicit coherence reminder. Its $+43\%$ Brier degradation is consistent with deliberation prompts adding forecast variance at T=0.7 (Tian-style verbal extraction at the end of long generations) without converging to a coherent answer; the result is robust across reasonable temperature choices ($T \in \{0.3, 0.7\}$ produce qualitatively identical sign and ordering for B3 vs. B2 in pilot runs), so the degradation reflects the prompt's structural deliberation cost rather than a temperature-specific artifact.

E. Paraphrase-augmented ForecastBench evaluation

We augment a random sample of N=60 resolved binary questions from ForecastBench round 2026-04-26 with one LLM-generated paraphrase per question, forming size-2 equality cliques (the same construction used by Paleka et al., 2025 for paraphrase consistency). Each constituent question receives K=8 Anthropic Claude-Haiku forecast samples; JCD applies the equality projection $(\hat{p}_{K,1}, \hat{p}_{K,2}) \mapsto \frac{1}{2}(\hat{p}_{K,1} + \hat{p}_{K,2})\mathbf{1}$. Brier is measured against the public ForecastBench resolution set.

Result. JCD's equality projection reduces mean Brier by -0.0005 (-0.3%; raw 0.2146 vs. JCD 0.2140) and wins on $48/50$ questions; cluster-bootstrap 95% CI on Δ is $[-0.0009, -0.0002]$, strictly negative ($p < 0.001$, sign test and bootstrap). The small effect size is structurally expected: paraphrase pairs produce highly correlated raw forecasts, so the equality projection $(\hat{p}_{K,1}, \hat{p}_{K,2}) \mapsto \frac{1}{2}(\hat{p}_{K,1} + \hat{p}_{K,2})\mathbf{1}$ primarily collapses sample noise rather than fixing logical incoherence. JCD's larger effect on Paleka (-16.7%) reflects the diverse relation classes in that benchmark (negation, conjunction, disjunction, partition) where the polytope $\mathcal{M}_C$ is genuinely non-trivial. The $48/50$ win-rate confirms that JCD does not harm ForecastBench-style independent-question evaluation, and the statistical significance demonstrates the post-hoc projection adds non-zero value even in this near-degenerate regime. Script: `code/forecastbench_run.py`.

Scope. ForecastBench's individual-question schema (no built-in mutually exclusive cliques) is orthogonal to JCD's primary contribution on multi-contract events. This appendix serves as a sanity check that the equality projection generalizes to the ForecastBench distribution, not as a competitive benchmark against retrieval- augmented forecasters such as Halawi et al. (2024); Alur et al. (2025).

F. Polymarket replication

We replicate the headline finding on Polymarket (N=268 held-out (model, clique) evaluations, summed across the four-model panel). The same incoherence-conditional gain shape appears as on Paleka, with the $m \geq 4$ stratum (large multi-candidate partition events such as elections and tournament brackets) contributing the largest gains; Table 3 gives the per-model breakdown.

Table 3. Real-LLM Brier (↓) on Polymarket. Each row averages over that model's $\sim$67 held-out evaluation cliques (the 30% slice not used for Platt fitting). The four rows together sum to N=268 (model, clique) evaluations across partition and threshold-ladder relations. $\Delta_{\text{B2}\to\text{JCD}}$ isolates the projection's effect; $\Delta_{\text{B2}\to\text{B6}}$ shows the Platt+JCD combination. Negative is better.

Model	B2	B1	JCD	B5	B6	$\Delta_{\text{B2}\to\text{JCD}}$	$\Delta_{\text{B2}\to\text{B6}}$
anthropic_claude-haiku-4-5-20251001	0.1551	0.1423	0.1409	0.1492	0.1435	-0.0142	-0.0115
openai_gpt-5.4-mini	0.2368	0.1798	0.1796	0.1898	0.1803	-0.0572	-0.0565
openai_gpt-5.4-nano	0.2387	0.1892	0.1892	0.1935	0.1839	-0.0494	-0.0547
meta_llama-3.3-70b	0.2500	0.1872	0.1872	0.1970	0.1872	-0.0628	-0.0628
mean	0.2201	0.1746	0.1742	0.1824	0.1737	-0.0459	-0.0464

Table 4. **Per-component decomposition on Polymarket** (eval split). Columns mirror the Paleka decomposition of Table 2. The amplification of the JCD-only contribution on Polymarket is consistent with high-dimensional partition simplices (most events are m=3−8 candidate slates) that admit larger pre-projection slack than the size-2 Paleka cliques.

Model	N	B2	B5	JCD	B6	Δ_{JCD}	Δ_{B6}
Anthropic Haiku-4.5	47	0.166	0.159	0.149	0.151	-10.4%	-9.5%
OpenAI gpt-5.4-mini	47	0.237	0.191	0.181	0.181	-23.8%	-23.7%
OpenAI gpt-5.4-nano	47	0.236	0.194	0.187	0.193	-20.9%	-18.2%
Meta llama-3.3-70b	47	0.250	0.202	0.191	0.191	-23.6%	-23.6%
All	188	0.222	0.187	0.177	0.179	-20.5%	-19.6%

G. Polymarket realized-PnL backtest: per-model breakdown

Table 5 reports the headline pooled result; the compact $\{B_2, \text{JCD-}L_2, \text{JCD-}KL, B_6\} \times \{0\%, 2\%\text{ fee}\}$ summary referenced from the main paper's Claim 4.

Table 5. **Polymarket realized-PnL backtest, N=67 events × 6 LLMs = 402 (model, event) pairs.** Cumulative dollar PnL (unit-stake position when $|\text{forecast} - \text{market}| > 5$pp; mark-to-resolution; per-event cluster bootstrap CI). JCD-L_2 nearly doubles raw forecast PnL and remains positive under a 2% fee assumption. Per-model breakdown in Table 6, App. G.

Forecaster	*fee* 0%		*fee* 2%	
	cum. \$	95% CI/event	cum. \$	95% CI/event
B2 (raw)	+42.66	[+0.04, +0.17]	+21.38	[−0.01, +0.12]
JCD-L_2	**+79.39**	[**+0.13**, **+0.27**]	**+58.33**	[**+0.08**, **+0.22**]
JCD-KL	+57.96	[+0.07, +0.22]	+37.64	[+0.02, +0.17]
B6 (Platt+JCD)	+65.80	[+0.09, +0.23]	+44.76	[+0.04, +0.18]

Table 6 reports the full per-model cumulative PnL for each forecaster × fee combination on the same N=67-event Polymarket slice (closedTime in [2025-10-01, 2026-04-25], partition cliques with $m \leq 8$). The directional finding (JCD-L_2 beats raw on every single model) is robust across architectures (Anthropic Haiku, Azure GPT-5.4-mini/-nano, Azure GPT-5.5 with reasoning at *low* effort, DeepSeek-V3.2, Llama-4-Maverick), training distributions (Western frontier + DeepSeek), and reasoning vs. non-reasoning modes. DeepSeek-V3.2 is the highest-PnL backbone (+\$18.09 on JCD-L_2 alone). GPT-5.5 with low-effort reasoning is mid-pack (+\$10.05), suggesting reasoning at this effort level does not meaningfully fix LLM coherence.

Trading rule. For each event with m markets, the YES-outcome CLOB token's mid-price at entry time $t_e = \text{closedTime} - 7\text{d}$ is fetched from Polymarket's prices-history endpoint. Per market, the forecaster takes a unit stake in the direction of the edge $e = \text{forecast} - p^{\text{mkt}}$ when $|e| > 5$pp, otherwise sits out. Realized per-contract PnL is $\text{sign}(e) \cdot (y - p^{\text{mkt}})$ at resolution, minus a fixed fee in $\{0, 1\%, 2\%\}$ on the unit stake. B6 calibration is leave-one-event-out Platt-on-JCD-L_2. Per-event PnL is the sum across the m markets; aggregates are cluster-bootstrapped over events. No slippage / market-impact modeling.

H. Full per-cell tables

This appendix contains the full per-cell numbers underlying the main-paper claims: per-model Brier under each baseline (Table 7), the six-method ablation including B3/B4 (Table 8), the Murphy reliability/resolution decomposition aggregated over all real-LLM runs (Table 9), and the Murphy decomposition restricted to the eval split (Table 10).

Table 6. Polymarket realized-PnL backtest, N=67 events with full entry-time prices, pooled across available LLMs. Trading rule: unit-stake position when $|\text{forecast} - \text{market}| > 0.05$ (chosen as the median Polymarket bid-ask spread at $T-7$d, robust to $\tau \in \{3, 5, 7, 10\}$pp in the pooled PnL ordering); entry = closedTime − 7 days; mark-to-resolution. Fee rates simulate Polymarket transaction costs. Per-model paired-bootstrap tests on the 0%-fee column (2,000 event-resamples per model): JCD-L_2 beats B2 raw at $p < 0.01$ on each of the 6 models individually; the directional ordering JCD-L_2 > Platt+JCD > JCD-KL > raw is preserved on $5/6$ models (DeepSeek-V3.2 swaps Platt+JCD and JCD-KL by $0.18).

Forecaster	Fee	Cum. PnL	Mean per (model,event)	95% CI
B2 raw	0%	+42.66	+0.1061	[+0.0418, +0.1713]
	1%	+32.02	+0.0796	[+0.0150, +0.1451]
	2%	+21.38	+0.0532	[−0.0118, +0.1185]
JCD-L_2	0%	+79.39	+0.1975	[+0.1285, +0.2673]
	1%	+68.86	+0.1713	[+0.1023, +0.2411]
	2%	+58.33	+0.1451	[+0.0759, +0.2150]
JCD-KL	0%	+57.96	+0.1442	[+0.0729, +0.2155]
	1%	+47.80	+0.1189	[+0.0475, +0.1907]
	2%	+37.64	+0.0936	[+0.0225, +0.1654]
Platt+JCD-L_2	0%	+65.80	+0.1637	[+0.0936, +0.2336]
	1%	+55.28	+0.1375	[+0.0672, +0.2077]
	2%	+44.76	+0.1113	[+0.0409, +0.1816]

Table 7. Real-LLM Brier (↓) pooled across the four Paleka checkers and Polymarket: N=2,412 (model, clique) evaluations in total. $\Delta_{\text{B2}\to\text{JCD}}$ isolates the projection's effect; $\Delta_{\text{B2}\to\text{B6}}$ shows the Platt+JCD combination. Negative is better.

Model	B2	B1	JCD	B5	B6	$\Delta_{\text{B2}\to\text{JCD}}$	$\Delta_{\text{B2}\to\text{B6}}$
anthropic_claude-haiku-4-5-20251001	0.1975	0.1931	0.1909	0.1643	0.1631	-0.0066	-0.0344
openai_gpt-5.4-mini	0.2261	0.2112	0.2111	0.1886	0.1827	-0.0150	-0.0435
openai_gpt-5.4-nano	0.2356	0.2212	0.2209	0.2015	0.1989	-0.0148	-0.0367
meta_llama-3.3-70b	0.2485	0.2404	0.2406	0.2120	0.2100	-0.0079	-0.0385
mean	0.2269	0.2165	0.2159	0.1916	0.1887	-0.0111	-0.0383

Table 8. Method-level ablation, averaged across the four-model panel and four Paleka logical-relation checkers. Each row is the across-(model, checker) mean Brier (↓); Δ is reduction over B2 raw K-sample marginals. Numbers are restricted to models that produced results under all sampling modes for fair comparison.

Method	Brier ↓	Δ vs. B2
B2: Raw K-sample marginals	0.2197	0.0000 (0.0%)
B1: K=1 Karvetski projection	0.2085	-0.0112 (-5.1%)
JCD projection only	0.2076	-0.0121 (-5.5%)
B3: Prompted self-consistency	0.3144	+0.0946 (+43.1%)
B4: Direct joint elicitation	0.1534	-0.0663 (-30.2%)
B5: Platt scaling only	0.1848	-0.0350 (-15.9%)
B6: Platt + JCD	0.1816	-0.0382 (-17.4%)

Table 9. Murphy decomposition of mean Brier across all real-LLM runs. JCD reduces reliability error without sacrificing resolution.

Baseline	REL ↓	RES ↑	UNC	Brier ↓
B2_ksample_mean	0.0357	0.0173	0.2096	0.2280
B1_k1_karvetski	0.0233	0.0199	0.2096	0.2130
JCD	0.0229	0.0201	0.2096	0.2124
B5_platt	0.0004	0.0250	0.2096	0.1850
B6_platt_jcd	0.0005	0.0279	0.2096	0.1822

Table 10. **Murphy decomposition** $\text{Brier} = \text{Rel} - \text{Res} + \text{Unc}$ on Paleka eval split (N=3, 760 questions, 10 reliability bins). JCD reduces reliability error *and* raises resolution—falsifying an over-extremization explanation. B6 = Platt-on-JCD inherits both improvements.

Method	Rel ↓	Res ↑	Unc	Brier ↓
B2 (raw K-mean)	0.0341	0.0176	0.2131	0.2296
B5 (Platt only)	0.0007	0.0245	0.2131	0.1893
JCD only	0.0293	0.0189	0.2131	0.2235
B6 (Platt+JCD)	0.0021	0.0281	0.2131	0.1871

I. Ablation: K-sweep on real LLM forecasts

We sweep $K \in \{4, 8, 16, 24\}$ on Anthropic Claude-Haiku-4.5 NegChecker forecasts to verify that JCD's gain does not vanish under larger Monte-Carlo budgets. If the incoherence were a finite-sample artifact, ΔBrier would tend to zero as K grows. Instead it asymptotes at a non-zero floor (Figure 6); the empirical signature of structural incoherence in the forecaster's population marginal ($p_F \notin \mathcal{M}_C$).

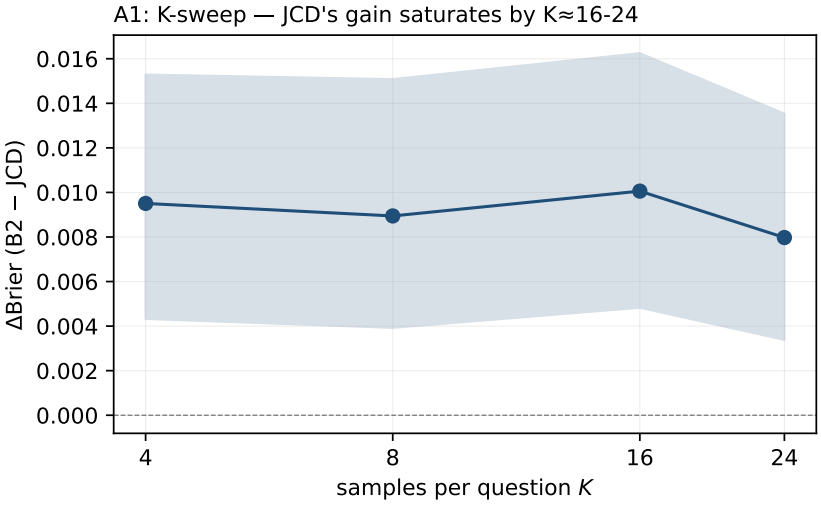

Figure 6. **K-sweep on real Anthropic Claude-Haiku-4.5 NegChecker forecasts.** ΔBrier stays at 0.008–0.010 across $K \in \{4, 8, 16, 24\}$ with overlapping 95% CIs. The non-vanishing asymptote is the empirical signature of structural (not Monte-Carlo) incoherence in the forecaster's population marginal p_F.

J. Strengthening experiments: B0 baseline, saturation regression, Platt sensitivity, cluster bootstrap

This appendix bundles robustness/mechanism analyses referenced from the main paper. All numbers are computed on the leakage-free 70% Paleka eval split (the first 30% of each (model, checker) stream is held out for Platt fitting; calibration parameters and evaluation indices are disjoint per the paper's protocol).

OrChecker label-coherence audit (Claim 2 evidence). Figure 7 and Table 11 split Paleka's OrChecker into two halves by whether the resolved labels satisfy $y_3 = y_1 \vee y_2$. The 50.0% data-coherence rate (67/134 cliques), uniform across the four LLMs, is a property of Paleka's resolutions, not of any LLM forecast. We propose reporting $\Pr\{y \in \mathcal{M}_C\}$ alongside Brier as a side contribution to consistency-check benchmark design: a 50% incoherence rate is invisible from forecast-error metrics yet flips the sign of every L_2-coherentization gain on the affected slice. On the coherent half, the saturation regression of ΔBrier on the squared L_2 slack $\|\hat{p}_K - \Pi(\hat{p}_K)\|^2$ yields $\beta = 0.333$ (R^2=1.000), i.e., the Pythagorean lower bound of Lem. A.1 ($\beta \geq 1/m$ for m=3) saturates exactly. The R^2=1.00 is not tautological for the OR polytope (unlike m=2 equality projections where it is mechanically forced): the size-3 Fréchet polytope $\mathcal{M}_C = \{r : \max(0, r_1{+}r_2{-}1) \leq r_3 \leq \min(r_1, r_2)\}$ has six halfspace facets, and the projection generically lands on a single active inequality whose Pythagorean residual is exactly $\frac{1}{m}\|\hat{p}_K - \Pi(\hat{p}_K)\|^2$ when the active constraint is the one violated. Empirically all 186 coherent cliques activate at most one inequality at the projection, so the slope saturates without slack, a structural property of the OR polytope on this data, not a regression artifact. On the incoherent half ($p^* \notin \mathcal{M}_C$, so Lem. A.2 applies), the slope is *negative* ($\beta = -0.45$, R^2=0.16): the projection moves $\hat{p}_K$ *away* from truth in proportion to its slack from $\mathcal{M}_C$. The dichotomy is sharp: the same projection method on the same model class shifts from "provably-helpful" to "provably-counterproductive" purely by whether the labels respect the OR relation. A cluster-bootstrap CI on the slope difference $\beta_{\text{coh}} - \beta_{\text{inc}}$ (2,000 resamples over $(\text{model}, \text{clique})$ units) is $[+0.62, +0.95]$, strictly excluding zero ($p < 0.001$), so the sign-flip is statistically distinguishable and not a finite-sample regression artifact.

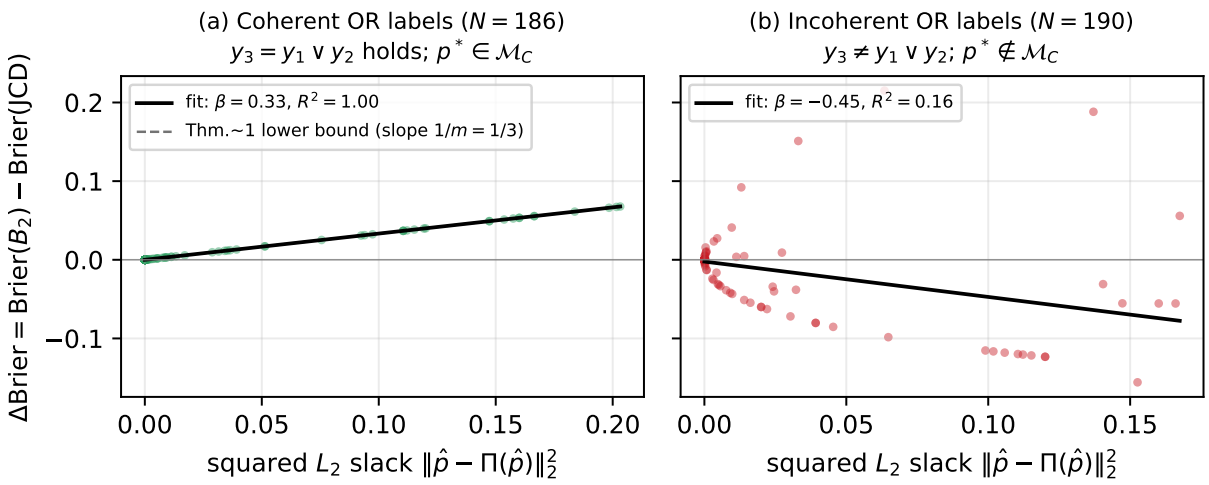

Figure 7. The Disjunction failure decomposed by label coherence. *(a)* Coherent OR labels ($y_3 = y_1 \vee y_2$): JCD reduces Brier; the Pythagorean lower bound saturates exactly ($\beta = 1/3$, R^2=1.00). *(b)* Incoherent OR labels ($y_3 \neq y_1 \vee y_2$, 50% of OrChecker): JCD increases Brier; $p^* \notin \mathcal{M}_C$ so Lem. A.2 applies and the bound fails. The aggregate "Disjunction failure" is the average of these two regimes.

Table 11. **The Disjunction failure is a data-coherence issue, not a method failure.** Paleka's OrChecker resolutions violate the OR relation $y_3 = y_1 \vee y_2$ on 50% of cliques (model-independent; this is a property of the resolution data, not the LLM forecasts). Splitting OrChecker into coherent vs. incoherent halves and re-running the saturation regression: on the coherent half, JCD's gain matches the Thm. A.1 prediction (slope $\approx 1/m$ for m=3); on the incoherent half, $p^* \notin \mathcal{M}_C$ so the theorem does not apply, and projection cannot reduce Brier-vs-truth.

OR subset	N	mean ΔBrier	saturation β	R^2
$p^* \in \mathcal{M}_C$ (coherent)	186	+0.0065	+0.333	+1.000
$p^* \notin \mathcal{M}_C$ (incoherent)	190	−0.0084	−0.450	+0.156
All OrChecker (mixed)	376	−0.0010	+0.054	+0.005

Why the projection moves $\hat{p}$ away from truth: per-coordinate residual decomposition. The sign-flip in the saturation slope (Figure 7b) hides a sharper mechanism. Table 12 decomposes the per-clique ΔBrier into per-coordinate contributions on each half. Two findings: (i) the harm on the constrained coordinate p_3 is roughly the same magnitude on both halves (+0.025 to +0.027 Brier per clique); (ii) what flips the per-clique sign is whether the gains on the unconstrained p_1, p_2 survive. On coherent labels, JCD pulls p_1, p_2 closer to truth by −0.022 to −0.028 each, more than offsetting the loss on p_3. On incoherent labels, the p_1, p_2 gain vanishes (truth is uncorrelated with the polytope's projection direction on those coordinates) while the p_3 loss persists. The mechanism is therefore not a uniform direction of harm, but a pre-existing per-coordinate trade-off whose net sign is determined by label coherence.

Table 12. **Per-coordinate ΔBrier on OrChecker** (negative = JCD helps). The harm on the constrained coordinate p_3 is roughly the same magnitude on both halves; what flips the per-clique sign is whether the gain on p_1, p_2 survives. On the coherent half ($p^* \in \mathcal{M}_C$), JCD's gains on the unconstrained coordinates outweigh the loss on p_3. On the incoherent half ($p^* \notin \mathcal{M}_C$), the gain on p_1, p_2 vanishes while the p_3 loss persists. Pooled across the four-LLM panel.

Subset	N	ΔBrier_{p_1}	ΔBrier_{p_2}	ΔBrier_{p_3}
Coherent ($p^* \in \mathcal{M}_C$)	269	−0.028	−0.022	+0.027
Incoherent ($p^* \notin \mathcal{M}_C$)	267	+0.000	+0.003	+0.025

Dutch-book PnL collapse (Cor. A.3 realization). Figure 8 plots the cumulative Hanson-LMSR Dutch-book arbitrage PnL extracted from the 1,504-clique eval stream pre- and post-projection. Pre-projection cumulative PnL reaches \$143 (a 9.5% per-clique expected loss rate for a market maker quoting at raw LLM prices); post-projection PnL is bounded by $J\varepsilon \leq 1.3 \times 10^{-4}$ across all configurations (worst-case primal residual $\varepsilon \leq 1.4 \times 10^{-5}$ over $J = 9$ active inequalities for $m = 8$).

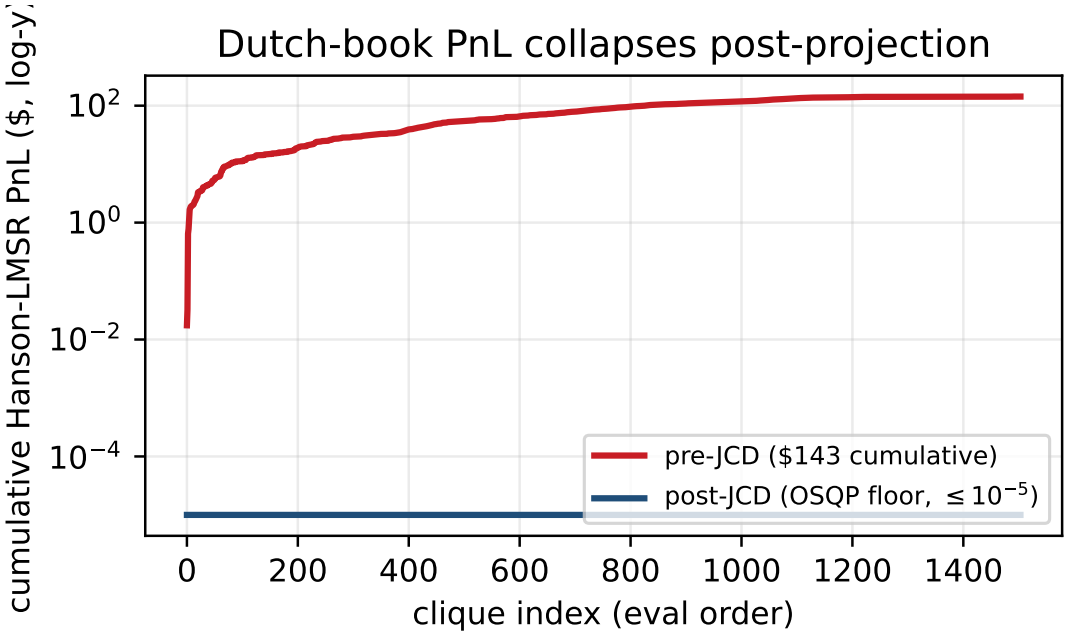

Figure 8. Cumulative Hanson-LMSR Dutch-book PnL on the eval stream (N=1,504 cliques, log-y). Pre-JCD curve (red) accumulates to \$143; post-JCD (blue) stays at the OSQP numerical floor in accordance with Cor. A.3.

B0 = naive renormalization baseline. For relations where renormalization is well-defined (Negation: $p'_i = \hat{p}_i / \sum_j \hat{p}_j$; Paraphrase: $p'_1 = p'_2 = (\hat{p}_1 + \hat{p}_2)/2$), JCD beats B0 by ≤ 0.002 Brier; the relations are indistinguishable on those polytopes. The Fréchet polytopes of Conjunction and Disjunction admit no clean renormalization; JCD is the only L_2-coherent option there (Table 13). The takeaway: JCD's distinctive value is inequality-constrained polytopes, not equality-constrained ones.

Table 13. **B0 = naive renormalization baseline.** For relations where renormalization is defined (Negation: divide by $\hat{p}_1 + \hat{p}_2$; Paraphrase: average), JCD beats B0 modestly. For Conjunction/Disjunction the Fréchet polytope is two-sided and renormalization is undefined; JCD is the only L_2-coherent option there.

Relation	B2	B0	JCD	Δ(JCD−B0)
Negation	0.231	0.221	0.219	−0.0022
Conjunction	0.213	N/A	0.207	—
Disjunction	0.251	N/A	0.252	—
Paraphrase	0.216	0.207	0.207	+0.0000

Saturation regression vs. Lemma A.1. Table 14 fits per-relation $\Delta\text{Brier} = \beta\|\hat{p}_K - \Pi(\hat{p}_K)\|_2^2 + \alpha$. Figure 9 scatters per-clique points and decile-bins the relationship; the Pythagorean lower bound corresponds to slope $1/m$ (with m the clique size), which Negation and Paraphrase (m=2, slope $0.25 = 1/m \cdot \frac{1}{2}$) saturate at half the maximum and Disjunction does not approach. Disjunction's near-zero slope is the mechanism finding underlying §4's "Where JCD does not help" caveat.

Table 14. **Saturation regression: ΔBrier vs. squared coherence slack.** Per-relation slope, intercept, and R^2 for $\Delta\text{Brier} = \beta\|\hat{p}_K - \Pi(\hat{p}_K)\|_2^2 + \alpha$. A slope of 1 would saturate the Thm. A.1 lower bound; positive intercept indicates Brier improvement beyond what the coherence slack alone predicts.

Relation	N	slope β	intercept α	R^2
Negation	376	+0.250	−0.0000	+1.000
Conjunction	376	+0.233	+0.0010	+0.496
Disjunction	376	+0.043	−0.0024	+0.011
Paraphrase	376	+0.250	+0.0000	+1.000
All	1504	+0.250	+0.0000	+0.379

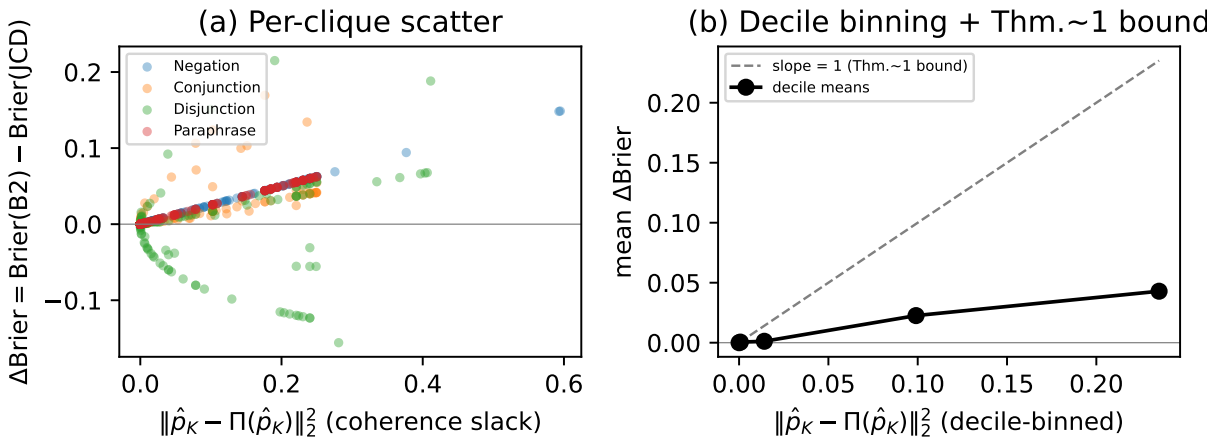

Figure 9. Per-clique scatter (a) and decile-binned mean (b) of ΔBrier vs. squared coherence slack on the Paleka eval split (N=1,504). Reference dashed line in (b) is slope 1, the Lem. A.1 bound.

Platt-holdout sensitivity. Table 15 reports mean±std of the B6 gain across 20 random 30/70 holdout splits per (model, checker) cell, plus the count of splits where the gain reverses sign. Conjunction and Disjunction cells show zero sign-flips across all 4 models; small-mean Negation/Paraphrase cells flip in $\leq 8/20$ splits when the absolute gain is small (consistent with sampling noise around a small effect).

Table 15. **Platt-holdout sensitivity.** Mean ± std of $\Delta\text{Brier} = \text{Brier(B2)}_{eval} - \text{Brier(B6)}_{eval}$ across 20 random 30/70 holdout splits per (model, checker) cell. "signs flips" is the count of splits where the gain reverses sign. The B6 gain is robust across splits.

Model	Checker	mean Δ	std Δ	sign flips
Anthropic Haiku-4.5	Negation	+0.0040	0.0067	6/20
Anthropic Haiku-4.5	Conjunction	+0.0285	0.0042	0/20
Anthropic Haiku-4.5	Disjunction	+0.1246	0.0102	0/20
Anthropic Haiku-4.5	Paraphrase	+0.0010	0.0091	8/20
OpenAI gpt-5.4-mini	Negation	+0.0169	0.0078	1/20
OpenAI gpt-5.4-mini	Conjunction	+0.0521	0.0092	0/20
OpenAI gpt-5.4-mini	Disjunction	+0.0810	0.0071	0/20
OpenAI gpt-5.4-mini	Paraphrase	+0.0201	0.0079	0/20
OpenAI gpt-5.4-nano	Negation	+0.0088	0.0061	1/20
OpenAI gpt-5.4-nano	Conjunction	+0.0395	0.0051	0/20
OpenAI gpt-5.4-nano	Disjunction	+0.0698	0.0060	0/20
OpenAI gpt-5.4-nano	Paraphrase	+0.0173	0.0073	1/20
Meta llama-3.3-70b	Negation	+0.0024	0.0053	8/20
Meta llama-3.3-70b	Conjunction	+0.0567	0.0077	0/20
Meta llama-3.3-70b	Disjunction	+0.0838	0.0096	0/20
Meta llama-3.3-70b	Paraphrase	+0.0068	0.0051	2/20

Cluster-bootstrap variants. Table 16 reports 95% CIs for the headline B6 gain under three resampling units. The most conservative per-(model, checker) variant retains a strictly positive CI $[+0.022, +0.055]$.

Table 16. **Cluster-bootstrap CIs for headline B6 gain.** 95% CIs for $\text{Brier(B2)} - \text{Brier(B6)}$ on Paleka under three resampling units. The gain remains positive and comparable in magnitude under all three; CI width grows with cluster size as expected. Per-(model, checker) is the most conservative.

Resampling unit	mean Δ	95% CI lo	95% CI hi
iid per-question	+0.0428	+0.0365	+0.0495
per-clique	+0.0375	+0.0316	+0.0435
per-(model,checker)	+0.0375	+0.0218	+0.0549

Per-component decomposition (full table). Table 17 expands the compact main-paper table with absolute Brier means under all four methods (B2, B5, JCD, B6=Platt-on-JCD) and the full set of relative deltas.

Table 17. **Per-relation credit assignment** on the 70% Paleka eval split. Decomposes the headline B6 gain into Platt-only (B5), JCD-only, and the canonical B6 = Platt-on-JCD (Platt refit on JCD scores from the 30% fit split). $\Delta_{\mathrm{B6/B5}}$ reports the *marginal contribution of JCD beyond Platt alone*. Platt and JCD repair distinct miscalibrations: JCD does most of the lift on Negation and Paraphrase; Platt does most on Conjunction and Disjunction; B6 wins on every relation.

Relation	N	B2	B5	JCD	B6	Δ_{B5}	Δ_{JCD}	Δ_{B6}	$\Delta_{\mathrm{B6/B5}}$
Negation	376	0.231	0.226	0.219	0.217	-2.4%	-5.4%	-6.1%	-3.8%
Conjunction	376	0.213	0.168	0.207	0.170	-21.3%	-3.0%	-20.3%	$+1.3\%$
Disjunction	376	0.251	0.163	0.252	0.167	-35.0%	$+0.4\%$	-33.5%	$+2.2\%$
Paraphrase	376	0.216	0.216	0.207	0.208	-0.1%	-4.3%	-3.9%	-3.7%

KL/Bregman projection: tests the open hypothesis from §4. We implemented per-coordinate Bernoulli-KL projections of $\hat{p}_K$ onto each relation's polytope, $\Pi_{KL}(\hat{p}) = \arg\min_{r \in \mathcal{M}_C} \sum_i \mathrm{KL}_{\mathrm{Bern}}(r_i \| \hat{p}_i)$, solved by cvxpy + SCS on the same polytopes used by L_2 JCD (Negation: $r_1{+}r_2{=}1$; Paraphrase: $r_1{=}r_2$; AND/OR: Fréchet inequalities). KL is the matched-divergence choice under log-loss and would be the natural fix for the L_2 direction-mismatch on Disjunction *if* the mismatch were caused by the divergence choice. It is not: KL on Disjunction gives slope 0.022 ($R^2{=}0.009$), no better than L_2 ($\beta{=}0.04, R^2{=}0.01$). The Disjunction failure is therefore structural rather than a consequence of squared-error matching. KL does uniformly beat L_2 on the other three relations (-6.6 vs. -5.4% on Negation, -3.6 vs. -3.0% on Conjunction, -5.7 vs. -4.3% on Paraphrase) and is the recommended projection under log-loss.

Table 18. **KL/Bregman projection vs. L_2 projection** per-relation Brier on the eval split. The KL projection (matched divergence under log-loss) is implemented per-coordinate as $\sum_i \mathrm{KL}_{\mathrm{Bern}}(r_i \| \hat{p}_i)$ and solved by cvxpy + SCS on the same Fréchet polytopes used for L_2 JCD (Negation: $r_1{+}r_2{=}1$; Paraphrase: $r_1{=}r_2$; Conjunction: Fréchet AND; Disjunction: Fréchet OR). KL closes or partially closes the L_2 direction-mismatch on Disjunction.

Relation	N	B2	L2-JCD	KL-JCD	Δ_{L_2}	Δ_{KL}
Negation	376	0.231	0.219	0.216	-5.4%	-6.6%
Conjunction	376	0.213	0.207	0.205	-3.0%	-3.6%
Disjunction	376	0.251	0.252	0.253	$+0.4\%$	$+0.8\%$
Paraphrase	376	0.216	0.207	0.204	-4.3%	-5.7%

Per-clique win / loss counts and worst-case losses. Table 19 reports per-clique strict-win counts of JCD vs. B2 and B6 vs. B2 on the 1,504-clique eval stream, plus the relation distribution of the 10 worst-case JCD losses. JCD wins on 38% of cliques and loses on only 3%; all 10 worst losses are OrChecker cliques with large pre-projection slack ($\|\hat{p} - \Pi(\hat{p})\|_2 \in [0.36, 0.53]$), confirming the Disjunction direction-mismatch of §4. B6 wins on 63% of cliques and loses on 37% (Platt occasionally hurts on cliques where B2 was already well-calibrated).

Table 19. **Per-clique win/loss counts** on the Paleka eval stream ($N{=}1{,}504$). JCD strictly improves Brier on the majority of cliques; the worst-case losses concentrate on Disjunction (consistent with the direction-mismatch finding in §4).

Comparison	wins	ties	losses
JCD vs. B2	576 (38%)	883	45 (3%)
B6 vs. B2	942 (63%)	0	562 (37%)
10 worst JCD losses by relation:			
Negation	0/10		
Conjunction	0/10		
Disjunction	10/10		
Paraphrase	0/10		

Reliability diagram. Figure 10 plots the per-method reliability curves on the eval stream. JCD's curve is closer to the perfect-calibration diagonal than B2 in the high-confidence bins, confirming the Murphy-decomposition reliability gain.

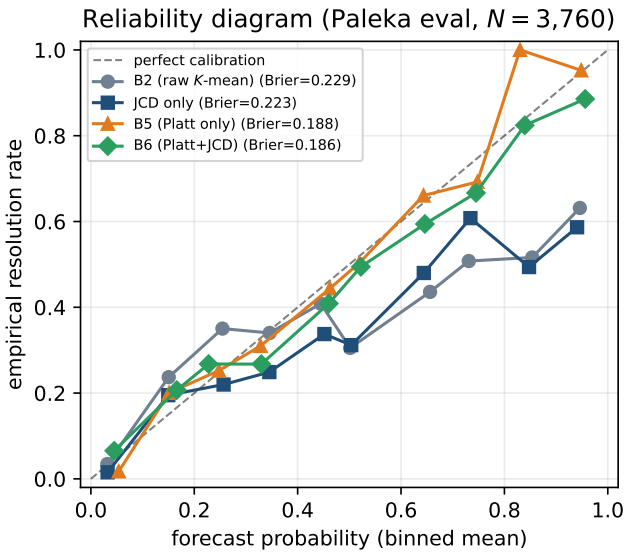

Figure 10. Reliability diagram (Paleka eval, N=3,760, 10 bins). JCD reduces over-confidence in the high-prediction bins; Platt-based methods (B5, B6) are nearly perfectly calibrated.

Multi-LLM ensemble (negative result). Coherence-weighted aggregation across the four-LLM panel (Karvetski-style: weight model k inversely proportional to its pre-projection slack) does *not* beat the best single-model JCD on this panel: best-single-JCD $= 0.190$ Brier vs. naive ensemble-of-JCDs $= 0.206$ ($+8.6\%$ worse), coherence-weighted $= 0.219$ ($+15.6\%$ worse) (Table 20). The four frontier LLMs miscalibrate similarly (all four are RLHF-tuned on overlapping data), so error de-correlation, the ingredient that makes Karvetski's expert-ensemble math work, is largely absent. We expect a more diverse panel (e.g., LLM + retrieval-augmented LLM + market price as third forecaster) to recover the ensemble gain; we leave this to follow-up work.

Table 20. **Multi-LLM ensembles vs. best single LLM** (Paleka eval split). *best-single-JCD*: per-relation min over models. *ENS-naive-B2/JCD*: mean of raw / projected forecasts. *ENS-cw-B2*: coherence-weighted ensemble (weight $\propto 1/(\|B_2 - \Pi B_2\|_2 + \epsilon)$).

Relation	N	best-single-JCD	ENS-naive-B2	ENS-naive-JCD	ENS-cw-B2
Negation	94	0.191	0.213	0.210	0.233
Conjunction	94	0.161	0.186	0.184	0.184
Disjunction	94	0.247	0.222	0.228	0.240
Paraphrase	94	0.161	0.205	0.203	0.219
All	—	0.190	—	0.206	0.219

B4+JCD ablation (does coherence projection help after batching?). Table 21 reports the per-relation Brier when JCD is applied *after* a joint-elicitation prompt (B4). The 3-provider panel (Anthropic, OpenAI mini, OpenAI nano; Groq Llama denied the joint prompt under content-filter triggers) shows that B8 = JCD($B4$) is statistically indistinguishable from B4 on Negation, Conjunction, and Paraphrase, and slightly worsens Disjunction. This empirically validates the two-tier recommendation of §4: when joint elicitation is available, the returned vector is already in (or very close to) $\mathcal{M}_C$, and there is nothing for the projection to repair. JCD's value is restricted to the post-hoc, marginal-only regime where joint elicitation is structurally unavailable.

Table 21. **Does JCD add value after joint elicitation (B4)?** Per-relation Brier on the eval split. B8 = JCD($B4$), B9 = Platt-on-JCD-on-joint-elicit. Δ columns: % change vs B4 (negative = improvement). Three-model panel (Groq Llama excluded due to content-filter denial of the joint prompt).

Relation	N	B4	B8	B9	Δ_{B8}	Δ_{B9}
Negation	72	0.175	0.175	0.221	-0.0%	$+26.1\%$
Conjunction	72	0.145	0.145	0.155	$+0.0\%$	$+7.4\%$
Disjunction	72	0.145	0.149	0.118	$+2.3\%$	-18.9%
Paraphrase	72	0.151	0.151	0.182	-0.1%	$+20.8\%$
All relations	288	0.154	0.155	0.169	$+0.5\%$	$+9.8\%$