FORMING PAIRWISE COHERENT EVENT PROBABILITIES FROM LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly asked for numerical probabilities of uncertain events. Individually, these estimates are often reasonable; collectively, they are frequently *incoherent*: no joint distribution realizes all of the reported marginal and pairwise-intersection probabilities at once. Existing remedies (prompting, calibration, fine-tuning, learned representations) reduce incoherence but do not eliminate it, while trivially coherent substitutes such as assuming independence discard the information in the judgments. We present, to our knowledge, the first model-agnostic post-processing method for LLM probabilities that *guarantees* exact pairwise coherence while certifiably staying closest to the original judgments. Starting from the classical fact that a co-occurrence matrix is pairwise coherent if and only if it is a probability mixture of binary outer products (i.e., lies in the full correlation polytope), we cast coherence enforcement as a projection onto this polytope up to a global rescaling, and solve it by column generation: an exactly solved 0-1 quadratic program selects new "possible worlds" and an active-set nonnegative least squares solver refits their weights. For every input and every stopping tolerance, the output is coherent by construction and comes with an explicit joint distribution certifying it; the stopping rule also certifies global optimality of the projection up to that tolerance. Deciding coherence is NP-complete, but hardness affects only worst-case running time, never correctness. On event families with known ground truth, four LLMs are substantially incoherent. Local Fréchet repairs and calibration-style rescaling (even when fit to the ground truth) reduce but never eliminate this incoherence, whereas our projection eliminates it, lowers root mean squared error in 35 of 40 model–premise settings, and attains the lowest error of all compared methods in 29.

1 INTRODUCTION

Large language models (LLMs) are routinely prompted for numerical probabilities of uncertain events, from factual claims to forecasts and outcomes of stochastic processes. In many applications these probabilities are used not one at a time but *jointly*: a forecaster prices several related markets, a decision-support system reasons about overlapping patient groups, an analyst conditions one event on another. Joint use requires more than individually accurate numbers. It requires *coherence*: a single probability distribution must reproduce all reported probabilities simultaneously. An incoherent collection admits a sure-loss ("Dutch book") betting strategy (de Finetti, 1974), and derived quantities such as conditional probabilities can become meaningless (e.g., exceed one).

LLM probability judgments are frequently incoherent (Zhu & Griffiths, 2024; Freedman & Toni, 2025; Paleka et al., 2025), and existing remedies *reduce* incoherence without eliminating it. Prompting and fine-tuning change the raw estimates but offer no guarantee; calibration maps (Platt et al., 1999; Guo et al., 2017) transform each probability separately and so cannot enforce a constraint that couples many probabilities; learned representations recover probabilities that are more, but not perfectly, coherent (Zhu et al., 2025); and local repairs that check each pair of events in isolation are insufficient because coherence is a *global* property (Figure 1). Conversely, coherent outputs are trivial to produce if fidelity is abandoned (e.g., by treating all events as independent), but this discards the dependence structure that the model did capture. No existing method for LLM probabilities delivers both exact coherence and fidelity to the model's judgments.

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

We close this gap for *pairwise coherence*: consistency of all marginal probabilities $P(E_i)$ and pairwise intersections $P(E_i \cap E_j)$ of a collection of events. It is classical that such a collection is coherent if and only if its co-occurrence matrix is a probability mixture of binary outer products $aa^\top$, $a \in \{0,1\}^N$, i.e., lies in the *full correlation polytope*, and that deciding membership in this polytope is NP-complete (Pitowsky, 1991; Deza & Laurent, 1997). One might conclude that exact coherence is out of reach and that heuristics must suffice. We show that this conclusion is mistaken: NP-hardness constrains the worst-case *running time* of an exact method, not the *correctness* of its output.

Our method projects an incoherent matrix onto the correlation polytope, up to a global rescaling that preserves all conditional probabilities. The projection is a nonnegative least squares (NNLS) problem over 2^N atoms, which we solve by column generation: an exactly solved 0-1 quadratic program identifies the "possible world" that best explains the current residual, and an active-set NNLS solver refits the mixture weights. Three properties make this an exact method rather than a heuristic (Section 4.1): (i) the output is coherent by construction for every input and every stopping tolerance, and is returned together with the joint distribution that realizes it; (ii) the pattern-selection step is solved to global optimality, so the method cannot stall in a poor local optimum; and (iii) the stopping tolerance only bounds the suboptimality of the projection, and this bound is certified at termination.

Contributions.

- **Exact coherence for LLM probabilities.** We give, to our knowledge, the first post-processing method for LLM probability judgments whose outputs are *guaranteed* to be pairwise coherent while certifiably remaining closest to the original judgments. It needs no model access, retraining, or ground truth, and applies on top of any elicitation strategy.
- **An exact, certifiable algorithm.** We reformulate coherence enforcement as an NNLS problem over the cone of the correlation polytope and prove that our column-generation solver always returns a coherent output, certifies global optimality of the projection up to a user-set tolerance, and is exact at zero tolerance. The characterization of coherence and the optimization primitives are classical; our contribution is the formulation that lets them deliver exact guarantees for this problem.
- **Empirical evidence.** On coin and dice event families with exact ground truth, four LLMs are substantially incoherent (normalized distance 0.1–0.4); a local Fréchet repair and a recalibration fitted with oracle access to the ground truth reduce but never remove this incoherence; and exact coherence lowers RMSE in 35 of 40 settings (average $0.307 \rightarrow 0.185$), the lowest of all compared methods.

2 Related Work

Coherence, the marginal problem, and correlation polytopes. Coherence goes back to de Finetti's Dutch book arguments (de Finetti, 1974) and is an instance of the *marginal problem*: deciding whether given low-order marginals are realizable by one joint distribution (Strassen, 1965). Tractable relaxations in graphical models enforce only local consistency (Wainwright & Jordan, 2008), and related consistency problems are intractable in quantum information (Liu et al., 2007). For binary variables with all first- and second-order moments specified, the realizable set is the full correlation polytope, whose membership problem is NP-complete (Pitowsky, 1991; Deza & Laurent, 1997). We use these classical results as our starting point and do not claim them as contributions.

Coherent adjustment of probability judgments. Closest to our goal is work on reconciling incoherent human judgments. The coherent approximation principle of Osherson & Vardi (2006) replaces a set of judgments by the closest coherent assignment; the authors show that this problem is NP-hard and search over probability distributions with simulated annealing, which yields coherent assignments but no guarantee that they are the closest ones. Predd et al. (2006) obtain scalability by projecting successively onto *local* coherence sets, explicitly relaxing global coherence. We instead focus on marginals and pairwise intersections, which lets us return a *globally* coherent output in every case *and* certify that it is (up to a tolerance) the closest one.

Probability judgments of LLMs. LLM probability judgments are systematically incoherent, in ways that resemble human probabilistic reasoning (Zhu & Griffiths, 2024; Freedman & Toni, 2025), and LLM forecasters fail arbitrage-based consistency checks (Paleka et al., 2025). Existing remedies reduce such violations without guaranteeing coherence: the ArbitrageForecaster of Paleka et al. (2025) improves consistency only on the checks it targets; Zhu et al. (2025) recover more coherent probabilities from model embeddings with an axiomatically constrained autoencoder, which requires access to internal representations; and chain-of-thought prompting (Wei et al., 2022) or preference fine-tuning (Ouyang et al., 2022) change the raw estimates but offer no guarantee. For binary (true/false) beliefs, Mitchell et al. (2022) enforce logical consistency post hoc with MaxSAT, which selects consistent truth values rather than coherent probabilities. Our method is complementary to all of these: it can be applied to whatever probabilities they produce and turns "more coherent" into "exactly coherent."

Calibration. Platt scaling (Platt et al., 1999), temperature and vector scaling (Guo et al., 2017), and related methods align predicted probabilities with observed frequencies (Dawid, 1982; Gneiting & Raftery, 2007). Calibration and coherence are distinct: calibration is a statistical property of individual predictions, whereas coherence is a joint feasibility constraint on a collection. Calibration maps act on each probability separately and provide no guarantee that the result is coherent, which we confirm with an oracle-fitted recalibration baseline (Section 5.2).

Matrix factorization and column generation. Writing a coherent matrix as $ADA^\top$ with binary A superficially resembles (sparse) nonnegative matrix factorization (NMF). The difference is essential: NMF learns continuous factors by local descent, without feasibility or optimality guarantees, whereas our atoms are the fixed, exponentially large set of binary outer products and new atoms are found by exact global optimization, which makes our output an explicit joint distribution with certifiable optimality. We solve the resulting problem with column generation (Lübbecke & Desrosiers, 2005) and active-set NNLS (Lawson & Hanson, 1995), classical tools that we use rather than claim.

3 Pairwise Coherence

3.1 Coherence Is a Global Property

Consider three events A, B, C with $P(A) = P(B) = P(C) = \frac{1}{2}$ and (Figure 1)

$$P(A \cap B) = \tfrac{1}{2}, \qquad P(B \cap C) = \tfrac{1}{2}, \qquad P(A \cap C) = 0. \tag{1}$$

Every pair is individually plausible: each intersection satisfies the Fréchet bounds $\max\{0, P(E_i) + P(E_j) - 1\} \le P(E_i \cap E_j) \le \min\{P(E_i), P(E_j)\}$. Yet the collection is infeasible: the first two equalities imply that A occurs if and only if B does, and B if and only if C does, so $P(A \cap C) = P(A) = \frac{1}{2} \neq 0$. Pairwise probabilities therefore cannot be validated, or repaired, one pair at a time: coherence constrains all events jointly.

3.2 Coherence as a Mixture of Possible Worlds

Let $E_1, \ldots, E_N$ be events on a probability space $(\Omega, \mathcal{F}, P)$ with indicators $X_i = \mathbf{1}(E_i)$. The *co-occurrence matrix* $J \in [0,1]^{N \times N}$ has entries

$$J_{ij} = P(E_i \cap E_j) = \mathbb{E}[X_i X_j], \tag{2}$$

so J is symmetric with the marginals $J_{ii} = P(E_i)$ on its diagonal. Each *outcome* $\omega \in \Omega$, i.e., one complete realization of the underlying random process (such as one particular sequence of coin flips), determines which of the N events occur. We record this in the binary event-membership vector $a(\omega) \in \{0,1\}^N$ with $a_i(\omega) = X_i(\omega)$ and call such a vector a *world*; for $N = 3$ events there are $2^3 = 8$ possible worlds. Writing d_a for the probability of world a,

$$J = \sum_{a \in \{0,1\}^N} d_a \, aa^\top, \qquad d_a \ge 0, \qquad \sum_a d_a = 1. \tag{3}$$

Conversely, every matrix of this form is the co-occurrence matrix of the events $\{a_i = 1\}$ under the distribution d over worlds. This yields the standard characterization on which we build.

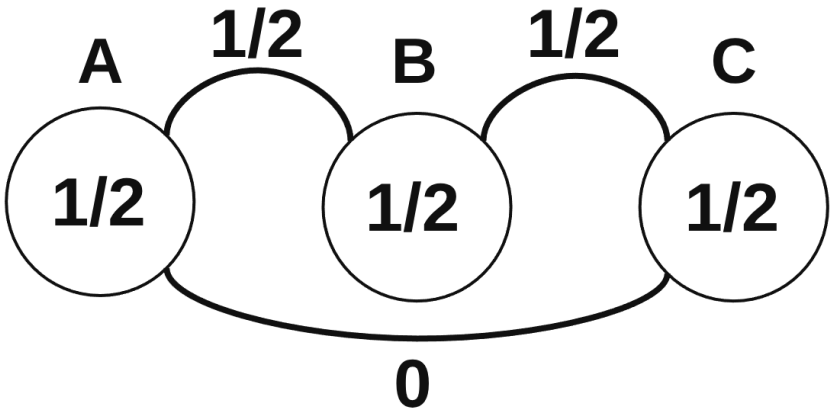

Figure 1: Pairwise assignments that are locally plausible (every pair satisfies its Fréchet bounds) but globally inconsistent. Circle labels are marginals; edge labels are pairwise intersections.

Definition 1 (Pairwise coherence)**.** *A symmetric matrix $J \in [0,1]^{N\times N}$ is* pairwise coherent *if there is a probability distribution $\{d_a\}_{a\in\{0,1\}^N}$ with $J = \sum_a d_a\, aa^\top$. Equivalently, $J = ADA^\top$ for a binary matrix A whose columns are worlds and a nonnegative diagonal matrix D with $\mathrm{tr}(D) = 1$ holding their probabilities.*

A distribution d satisfying Definition 1 is a *certificate* of coherence: given d, anyone can verify coherence by evaluating (3).

Scope. Pairwise coherence guarantees that all marginals and pairwise intersections are realized by one joint distribution. It does not determine higher-order quantities such as $P(E_i \cap E_j \cap E_k)$, which may differ across distributions that share the same J.

Complexity. The set $\mathcal{C}$ of pairwise-coherent matrices is the convex hull of $\{aa^\top : a \in \{0,1\}^N\}$, the *full correlation polytope*. Deciding membership in $\mathcal{C}$ is NP-complete (Pitowsky, 1991), so no algorithm can decide the coherence of every input in polynomial time unless P = NP. Section 4 shows how to nonetheless *enforce* coherence exactly.

4 Exact Projection onto the Coherent Set

Given an LLM-derived co-occurrence matrix J, which is in general incoherent, we seek a coherent matrix $\hat{J}$ that stays as close as possible to J. We require that (R1) $\hat{J}$ is *exactly* coherent for every input, certified by an explicit joint distribution, and (R2) $\hat{J}$ is optimally close to J, with a certificate of optimality. Because J is coherent if and only if it equals its own projection, computing the projection exactly is NP-hard in general. Rather than relaxing (R1) or (R2), as heuristic or local methods do, we satisfy both and accept exponential running time in the worst case, just as exact integer-programming solvers do. Section 5.5 shows that this worst case did not arise on the LLM outputs we study.

Projection problem. Let $\mathcal{K} = \{tP : t \geq 0,\ P \in \mathcal{C}\} = \{\sum_a w_a\, aa^\top : w \geq 0\}$ be the conic hull of $\mathcal{C}$. We define the projection of J as

$$\min_{t\geq 0,\ P\in\mathcal{C}} \ \|J - tP\|_F^2\,, \tag{4}$$

i.e., the coherent matrix P that, up to a single global scale factor t, best reproduces J. The scale factor makes the distance insensitive to a uniform over- or under-estimation of all probabilities: rescaling preserves every ratio between entries of J, and in particular every conditional probability $P(E_j \mid E_i) = J_{ij}/J_{ii}$. Substituting $M = tP$ shows that (4) is equivalent to the NNLS problem

$$\min_{w\geq 0} \ \Big\|J - \sum_{a\in\{0,1\}^N} w_a\, aa^\top\Big\|_F^2, \tag{5}$$

in which the weights need not sum to one. Given a solution $M = \sum_a w_a\, aa^\top$ of (5) with total weight $S = \sum_a w_a$, we return

$$\hat{J} = M/\max\{1, S\}, \tag{6}$$

which is an optimal P in (4) with $t = \max\{1, S\}$. When $S \le 1$ no rescaling is performed: $\hat{J} = M$ then lies in $\mathcal{C}$ (the remaining mass $1 - S$ is placed on the all-zero world), and since $\mathcal{C} \subset \mathcal{K}$, $\hat{J}$ is then also the Euclidean projection of J onto $\mathcal{C}$ (up to the tolerance in Proposition 2).[1]

Sparsity. Although problem (5) has 2^N atoms, by Carathéodory's theorem for cones it always has an optimal solution supported on at most $N(N+1)/2$ worlds, the dimension of the space of symmetric $N \times N$ matrices. We therefore use column generation (Lübbecke & Desrosiers, 2005): we maintain a small active set $\mathcal{A}$ of worlds and add a world only when it can improve the fit.

Pricing: selecting a new world. Let M be the current fit and $R = J - M$ the residual. The partial derivative of the objective of (5) with respect to w_a is $-2\, a^\top R a$, so world a can decrease the objective if and only if $a^\top R a > 0$. We therefore solve

$$g^\star = \max_{a \in \{0,1\}^N,\ a \notin \mathcal{A}} a^\top R a, \tag{7}$$

an unconstrained binary quadratic program, *to global optimality* by branch-and-bound. If $g^\star < \tau$ we stop; otherwise we add the maximizer to $\mathcal{A}$. Intuitively, (7) finds the world whose co-occurrence pattern best explains the remaining error.

Refitting weights. Given $\mathcal{A}$, we refit the weights by solving the restricted problem $\min_{w \ge 0} \|J - \sum_{a \in \mathcal{A}} w_a\, aa^\top\|_F^2$ with the active-set method of Lawson & Hanson (1995), which removes worlds whose weights reach zero. After each refit every retained world has positive weight and satisfies $a^\top R a = 0$. Algorithm 1 (Appendix A) gives the full procedure; it is precisely the Lawson–Hanson method applied to (5), with its column-selection step carried out implicitly by (7).

4.1 Guarantees: What Is Exact

Proposition 1 (Exact coherence)**.** *For every input J and every $\tau > 0$, the output $\hat{J}$ of Algorithm 1 is pairwise coherent. With w the final weights on $\mathcal{A}$ and $S = \sum_{a \in \mathcal{A}} w_a$, the distribution $d_a = w_a/\max\{1, S\}$ for $a \in \mathcal{A}$, together with $d_{\mathbf{0}} = 1 - S/\max\{1, S\}$ on the all-zero world, satisfies $d \ge 0$, $\sum_a d_a = 1$, and $\hat{J} = \sum_a d_a\, aa^\top$.*

Proposition 2 (Certified global optimality)**.** *Let M be the fit at termination, so that $\hat{J} = M/\max\{1, S\}$, and let $M^\star$ be an optimal solution of (5). Then*

$$0 \;\le\; \|J - M\|_F^2 - \|J - M^\star\|_F^2 \;\le\; 2\tau\,\mathrm{tr}(M^\star) \;\le\; 2\tau\sqrt{N}\,\|J\|_F, \tag{8}$$

and $\|M - M^\star\|_F^2 \le 2\tau\,\mathrm{tr}(M^\star)$. Hence $M \to M^\star$ as $\tau \to 0$, and with the exact stopping test $g^\star \le 0$ the algorithm terminates finitely at an exact optimum of (5). If $S \le 1$, then $\|J - \hat{J}\|_F^2 \le \min_{P \in \mathcal{C}} \|J - P\|_F^2 + 2\tau\,\mathrm{tr}(M^\star)$.

Proofs are given in Appendix B. Together, the two propositions make precise the sense in which our method is exact rather than heuristic: (i) *Coherence is exact.* It is a hard constraint that holds for every input and every τ, not an approximation that improves as τ shrinks; $\hat{J}$ is returned with the joint distribution that realizes it. (ii) *Pattern selection is exact.* Problem (7) is solved to global optimality, so the method cannot stall in a poor local optimum, and the termination test certifies that $a^\top R a < \tau$ for *all* 2^N worlds. (iii) *τ is a numerical tolerance.* Like the tolerance of any convex solver, τ bounds the suboptimality of the projection through (8); it never affects coherence. Empirically, results are insensitive to τ (Section 5.4). NP-hardness enters only through the worst-case running time of (7).

[1] Column generation applies equally to the Euclidean projection onto $\mathcal{C}$ by adding the constraint $\sum_a w_a \le 1$ to (5); we use (4) because it preserves conditional probabilities.

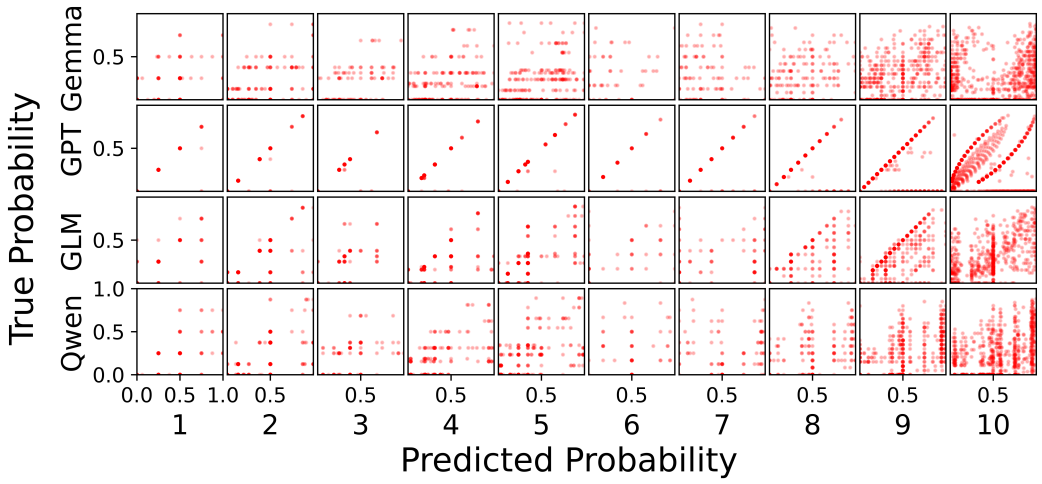

Figure 2: Raw predicted vs. true probabilities for coin- and dice-based statements. Rows: LLM. Columns: premise. Each dot is one statement.

Table 1: Premises defining the probability spaces. Dice premises read a `6-sided dice`, b `8-sided dice`, c `12-sided dice`, d `20-sided dice are rolled.` with the listed counts (a, b, c, d).

No.	Coin premise	No.	Dice counts (a, b, c, d)
1	`A fair coin is flipped 2 times.`	6	$(1, 0, 0, 0)$
2	`A fair coin is flipped 3 times.`	7	$(0, 1, 0, 0)$
3	`A fair coin is flipped 4 times.`	8	$(0, 0, 1, 0)$
4	`A fair coin is flipped 5 times.`	9	$(0, 0, 0, 1)$
5	`A fair coin is flipped 6 times.`	10	$(1, 1, 1, 1)$

Implementation and cost. We solve problem (7) with SCIP (MIT license) through PySCIPOpt (Maher et al., 2016), adding constraints that exclude worlds already in $\mathcal{A}$, and set $\tau = 10^{-4}$. The refit is a convex NNLS problem over $|\mathcal{A}|$ worlds and is inexpensive. The pricing problem (7) contains max-cut as a special case, so its worst-case cost is exponential in N and it dominates the running time; in practice, branch-and-bound exploits the structure of the residual, and projection of LLM outputs is far faster than projection of unstructured matrices of the same size (Section 5.5).

5 Experiments

We evaluate on probabilistic statements whose ground-truth probabilities are known exactly, and ask: (1) how accurate LLM probability estimates are, (2) how coherent they are as a collection, (3) whether enforcing exact coherence improves their accuracy, and (4) whether exact projection is fast enough in practice.

5.1 Experimental Setup

Event families. Following Zhu et al. (2025), we construct event families from simple stochastic processes. Each family is defined by a *premise* that specifies a probability space, such as repeated coin flips or dice rolls (Table 1). From each premise we generate natural-language *base events*, such as `the coin lands on heads exactly (X) times` or `the coin lands on heads more than (X) times`, and discard those with true probability below 0.1 or above 0.9 to avoid trivial cases. From the remaining base events $E_1, \ldots, E_N$ we construct (i) single-event

Table 2: Normalized distance to the pairwise-coherent set (lower is better) for raw LLM outputs (Raw), after Fréchet repair (Fr.), and after oracle log-probability scaling (Sc.), for premises 1–8. Our projection attains distance exactly 0 in every setting (Proposition 1) and is therefore omitted.

	Gemma 2			GPT-OSS			GLM-4.5			Qwen 2.5		
Premise	Raw	Fr.	Sc.	Raw	Fr.	Sc.	Raw	Fr.	Sc.	Raw	Fr.	Sc.
1	.264	.125	.144	.138	.117	.133	.192	.128	.124	.253	.147	.107
2	.353	.120	.180	.130	.102	.129	.222	.111	.126	.279	.122	.123
3	.224	.127	.098	.138	.114	.136	.142	.120	.082	.190	.101	.080
4	.276	.115	.131	.097	.096	.097	.234	.111	.096	.275	.092	.069
5	.317	.111	.167	.145	.102	.139	.224	.105	.086	.278	.094	.072
6	.210	.109	.129	.216	.154	.216	.291	.169	.126	.129	.125	.091
7	.232	.082	.144	.214	.141	.214	.359	.165	.174	.377	.146	.154
8	.279	.125	.082	.261	.133	.257	.280	.144	.152	.239	.120	.097

statements (premise + E_i) and (ii) pairwise conjunctions (premise + E_i + E_j), for a total of 2420 statements across all premises.

LLM elicitation. Each statement is queried independently, prefixed by the standardized instruction[2]

```
The following event has what probability? Respond only with
a number between 0 and 1 inclusive (no text).
```

We evaluate four models (Hugging Face identifiers): `google/gemma-2-2b-it` (Gemma license), `openai/gpt-oss-120b` (Apache 2.0), `zai-org/GLM-4.5` (MIT), and `Qwen/Qwen2.5-7B-Instruct-1M` (Apache 2.0). For each model and premise, the answers form a co-occurrence matrix J with J_{ii} the predicted probability of E_i and J_{ij} that of $E_i \cap E_j$. We deliberately use one standardized zero-shot protocol; our method is agnostic to how J is elicited (Section 6).

Baselines. We compare our projection with four alternatives. **Raw**: the unmodified LLM outputs. **Fréchet repair**: each off-diagonal entry is clipped to its Fréchet bounds $[\max\{0, J_{ii} + J_{jj} - 1\},\ \min\{J_{ii}, J_{jj}\}]$, a local correction that does not enforce coherence. **Scaling**: an affine recalibration of log-probabilities, $\log p \mapsto \alpha \log p + \beta$, in the spirit of Platt, temperature, and vector scaling (Platt et al., 1999; Guo et al., 2017), with α, β fitted to the *ground-truth* log-probabilities; this baseline thus has oracle access to labels that our method never sees, and its output is not guaranteed to be coherent. **Independence**: a coherent baseline that ignores dependence, $\hat{J}_{ii} = \max_j J_{ij}$ and $\hat{J}_{ij} = \hat{J}_{ii}\hat{J}_{jj}$ for $i \neq j$, i.e., independent events whose marginals are raised so that no intersection exceeds its marginal. Our projection uses neither the premise nor any ground truth.

5.2 LLM Judgments Are Informative but Systematically Incoherent

Raw predictions correlate positively with the true probabilities across models and premises (Figure 2), but errors are substantial and grow with the complexity of the premise (more coin flips or more dice). To measure joint (in)coherence, for each model and premise we compute the normalized distance $\mathrm{Dist}(J) = \|J - \hat{J}\|_F / N \in [0, 1]$, i.e., the root-mean-square difference over all N^2 entries between J and the coherent matrix $\hat{J}$ returned by our projection. As a sanity check, true co-occurrence matrices have distance 0 up to numerical precision.

Table 2 reports distances for premises 1–8. Evaluating the distance of each raw, Fréchet-repaired, and rescaled matrix requires a separate projection per matrix, which becomes expensive for premises 9–10, whose many base events yield the largest matrices; RMSE (Section 5.3) is cheap to compute and is reported for all ten premises. Incoherence is substantial for every model and premise evaluated, with raw distances typically between 0.1 and 0.4. The inconsistencies are therefore not isolated errors but systematic violations of joint feasibility. Both corrections reduce incoherence, yet

[2]Reported verbatim, including its slightly unusual wording, for reproducibility.

Table 3: RMSE between predicted and true probabilities (lower is better). Ind.: independence baseline; Fr.: Fréchet repair; Sc.: oracle log-probability scaling. Only Ind. and Ours are guaranteed to return coherent outputs. Bold: lowest error per model and premise.

	Gemma 2					GPT-OSS					GLM-4.5					Qwen 2.5				
Prem.	Raw	Ind.	Fr.	Sc.	Ours	Raw	Ind.	Fr.	Sc.	Ours	Raw	Ind.	Fr.	Sc.	Ours	Raw	Ind.	Fr.	Sc.	Ours
1	.364	.640	.269	.313	**.188**	.110	.227	.134	**.110**	.120	.216	.338	.173	.195	**.154**	.380	.653	.276	.299	**.190**
2	.390	.754	.254	.301	**.158**	**.099**	.190	.115	**.099**	.118	.226	.464	.145	.198	**.143**	.389	.700	.225	.301	**.184**
3	.358	.541	.275	.291	**.187**	**.115**	.227	.131	**.115**	.118	.213	.291	.196	.187	**.139**	.318	.588	.212	.252	**.183**
4	.368	.631	.251	.289	**.166**	**.069**	.153	.108	**.069**	.090	.248	.533	.154	.182	**.128**	.300	.734	.166	.196	**.134**
5	.399	.704	.272	.317	**.186**	.073	.201	.112	**.073**	.135	.230	.458	.169	.184	**.144**	.300	.640	.183	.201	**.165**
6	.341	.433	.288	.288	**.243**	.240	.446	.193	.240	**.177**	.341	.629	.337	.278	**.213**	.407	.422	**.264**	.310	.367
7	.402	.586	.282	.320	**.272**	.226	.441	**.176**	.226	.179	.419	.637	.408	.304	**.226**	.496	.742	.321	.366	**.225**
8	.320	.555	.323	.252	**.211**	.300	.524	**.171**	.300	.199	.329	.498	.230	.277	**.169**	.405	.609	.233	.262	**.211**
9	.333	.670	.302	.254	**.237**	.302	.598	**.149**	.302	.228	.233	.498	**.157**	.202	.198	.412	.707	.231	.253	**.180**
10	.488	.764	.321	.332	**.211**	.395	.678	**.230**	.379	.235	.290	.587	.244	.214	**.206**	.440	.736	.241	.297	**.191**
Mean	.376	.628	.284	.296	**.206**	.193	.368	**.152**	.191	.160	.274	.493	.221	.222	**.172**	.385	.653	.235	.274	**.203**

neither reaches zero in any setting: Fréchet repair fixes each pair in isolation, which cannot resolve violations that involve several events at once (Figure 1), and scaling transforms each probability separately, even though it is fitted to the ground truth. Our projection, in contrast, is coherent by construction.

5.3 Exact Coherence Improves Accuracy

For each model and premise, we measure accuracy by the root mean squared error between predicted and true probabilities over all n statements (single events and pairwise conjunctions), $\mathrm{RMSE} = \left(\frac{1}{n}\sum_{s=1}^{n}(\hat{p}_s - p_s)^2\right)^{1/2}$ (Table 3). Our projection lowers RMSE relative to the raw outputs in 35 of 40 settings, including all 30 settings for Gemma 2, GLM-4.5, and Qwen 2.5, and reduces the average RMSE from 0.307 to 0.185. It attains the lowest error of all five methods in 29 of 40 settings and outperforms Fréchet repair in 32 and oracle scaling in 34, even though neither of those baselines is required to produce a coherent output. The only settings where projection increases error are GPT-OSS on the five coin premises, whose raw estimates are already the most accurate in the study (RMSE ≤ 0.115); on the dice premises GPT-OSS benefits substantially (e.g., premise 8: $0.300 \to 0.199$). The coherent independence baseline increases error in all 40 settings, so coherence alone is not enough: it must be enforced while preserving the dependence structure in J, which our projection does.

Why coherence improves accuracy. The improvement stems from coherence itself rather than from knowledge of the task. The projection sees neither the premise nor any ground truth; the only structure it imposes is coherence, which the true matrix $J^\star$ satisfies. Because $J^\star$ lies in the closed convex cone $\mathcal{K}$, the Euclidean projection $M^\star$ of J onto $\mathcal{K}$ satisfies $\|M^\star - J^\star\|_F \leq \|J - J^\star\|_F$, a standard property of projections onto convex sets. Projecting onto $\mathcal{K}$ can thus only move the estimate closer to the truth in Frobenius norm (up to the tolerance of Proposition 2); the final rescaling (6) and the different weighting of entries in the RMSE mean that gains in RMSE are typical but not guaranteed, as the GPT-OSS coin settings show.

5.4 Sensitivity to the Tolerance τ

Table 4 (Appendix C) reports RMSE and runtime for $\tau \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ on premises 1–8. Across this entire range, RMSE changes by at most 5×10^{-4} in any setting and is identical to four decimals for all $\tau \leq 10^{-3}$, while runtime grows by 23% on average from $\tau = 10^{-2}$ to 10^{-5} and saturates below 10^{-4}. Coherence is exact at every τ (Proposition 1); τ only trades a small amount of computation for a tighter certified fit.

5.5 Exact Projection Is Fast on LLM Outputs

Figure 3 compares projection runtimes (on a 5.7 GHz CPU) for matrices derived from LLM predictions and for random matrices of the same size. Projection is markedly faster on LLM outputs,

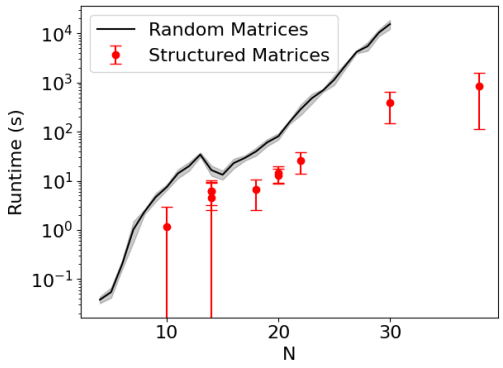

Figure 3: Projection runtime for co-occurrence matrices produced by LLMs (red) and for unstructured random matrices $(A + A^\top)/2$ with $A_{ij} \sim \mathcal{U}[0, 1]$ of the same size (black). N is the number of events in a premise. Margins show standard deviations over the four LLMs (structured) or five random draws (random).

taking between 3 s and 3 min on premises 1–8 (Table 4) and longer on the largest premises (Figure 3). LLM probability estimates thus carry exploitable structure: a small set of worlds explains them well, which branch-and-bound finds quickly. The worst-case exponential cost implied by NP-hardness did not materialize in our experiments.

6 Discussion

Collections of LLM probability judgments frequently fail to correspond to any joint distribution; this failure can be removed *exactly*, post hoc, and without model access, and doing so improves accuracy in most settings because the truth itself is coherent.

Relation to better elicitation. Chain-of-thought, few-shot, or joint prompting may reduce incoherence, and quantifying how much is a valuable direction. Such strategies are complementary rather than alternatives to our method: they change the raw estimates but cannot certify that a collection is coherent, and even a prompt that is coherent on average may occasionally return an infeasible collection. Our projection can be applied to the output of any elicitation strategy and turns "usually coherent" into "always coherent."

Limitations. First, we enforce coherence of marginals and pairwise intersections only; extending exact guarantees to higher-order intersections is possible in principle but increases the combinatorial cost. Second, our measurements use a single standardized zero-shot prompt and one query per statement, so the *magnitude* of incoherence may vary with prompting and sampling; whether incoherence stems from systematic bias or from sampling variability, the resulting collection is equally infeasible, and our guarantee applies either way. Third, we evaluate on synthetic domains, chosen because they provide exact ground truth; evaluating real-world forecasting domains is an important next step. Fourth, projection modifies the model's outputs; the global rescaling in (4) preserves all conditional probabilities, but whether the projected distribution reflects the model's implicit beliefs remains open. Finally, worst-case running time is exponential in N, although we did not encounter this regime on LLM outputs.

Implications. Exact coherence matters wherever related probabilities are used jointly: forecasts across related markets, where incoherence creates arbitrage opportunities, or risk estimates across overlapping patient groups, where it can lead to inconsistent treatment of populations. Our results point toward a hybrid approach in which LLMs provide flexible local estimates and a certified structured layer guarantees their global consistency.

AI USE STATEMENT

In this work, we used generative AI tools for writing, finding related work, developing research ideas, code, and proofs. We have not used generative AI tools for generating synthetic data. We have reviewed all AI-assisted work. All LLM generated text and code was carefully reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Bruno de Finetti. *Theory of Probability*. Wiley, 1974.

Michel M Deza and Monique Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.

Gabriel Freedman and Francesca Toni. Exploring the potential for large language models to demonstrate rational probabilistic beliefs. *arXiv preprint*, 2025.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *ICML*, 2017.

Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

Yi-Kai Liu, Matthias Christandl, and Frank Verstraete. Quantum computational complexity of the n-representability problem. *Physical Review Letters*, 98(11):110503, 2007.

Marco E Lübbecke and Jacques Desrosiers. Selected topics in column generation. *Operations research*, 53(6):1007–1023, 2005.

Stephen Maher, Matthias Miltenberger, João Pedro Pedroso, Daniel Rehfeldt, Robert Schwarz, and Felipe Serrano. PySCIPOpt: Mathematical programming in python with the SCIP optimization suite. In *Mathematical Software – ICMS 2016*, pp. 301–307. Springer International Publishing, 2016. doi: 10.1007/978-3-319-42432-3_37.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *EMNLP*, pp. 1754–1768, 2022.

Daniel Osherson and Moshe Y Vardi. Aggregating disparate estimates of chance. *Games and Economic Behavior*, 56(1):148–173, 2006.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. *ICLR*, 2025.

Itamar Pitowsky. Correlation polytopes: their geometry and complexity. *Mathematical Programming*, 50(1):395–414, 1991.

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

Joel B Predd, Sanjeev R Kulkarni, H Vincent Poor, and Daniel N Osherson. Scalable algorithms for aggregating disparate forecasts of probability. In *2006 9th International Conference on Information Fusion*, pp. 1–8. IEEE, 2006.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Jian-Qiao Zhu and Thomas L Griffiths. Incoherent probability judgments in large language models. *arXiv preprint*, 2024.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Recovering event probabilities from large language model embeddings via axiomatic constraints. *arXiv preprint*, 2025.

A Algorithm

Algorithm 1 is the Lawson–Hanson active-set method (Lawson & Hanson, 1995) for the NNLS problem (5), in which the column with the largest gradient is found implicitly by solving the binary quadratic program (7) rather than by enumerating all 2^N columns.

Algorithm 1 Exact projection onto pairwise-coherent matrices

Input: symmetric $J \in [0,1]^{N\times N}$, tolerance $\tau > 0$
Initialize active set $\mathcal{A} \leftarrow \emptyset$, weights $w^* \leftarrow 0$, fit $M \leftarrow 0$, residual $R \leftarrow J$
while true do
  (Pricing) Solve to global optimality (0-1 quadratic program):
    $a^\star \in \arg\max_{a\in\{0,1\}^N,\ a\notin\mathcal{A}} a^\top R a, \quad g^\star \leftarrow a^{\star\top} R a^\star$
  if $g^\star < \tau$ **then**
    break *(certifies $a^\top R a < \tau$ for all $a \in \{0,1\}^N$)*
  end if
  Add world: $\mathcal{A} \leftarrow \mathcal{A} \cup \{a^\star\}$, with $w^*_{a^\star} \leftarrow 0$
  (Refit) Solve the unconstrained least squares problem over $\mathcal{A}$:
    $w \in \arg\min_{u\in\mathbb{R}^{|\mathcal{A}|}} \left\| J - \sum_{a\in\mathcal{A}} u_a\, aa^\top \right\|_F^2$
  while $\exists\, a \in \mathcal{A}$ s.t. $w_a \le 0$ **do**
    $\mathcal{N} \leftarrow \{a \in \mathcal{A} : w_a \le 0\}$ *(violating worlds)*
    Maximal feasible step toward w: $\alpha \leftarrow \min_{a\in\mathcal{N}} \dfrac{w^*_a}{w^*_a - w_a}$
    Update feasible iterate: $w^* \leftarrow w^* + \alpha\,(w - w^*)$
    Remove all worlds with $w^*_a = 0$ from $\mathcal{A}$ (this includes the blocking world attaining the minimum in α)
    Re-solve the unconstrained least squares problem over the reduced $\mathcal{A}$ to obtain w
  end while
  $w^* \leftarrow w$; $\quad M \leftarrow \sum_{a\in\mathcal{A}} w^*_a\, aa^\top$; $\quad R \leftarrow J - M$ *(now $a^\top R a = 0$ for all $a \in \mathcal{A}$)*
end while
(Normalize) $S \leftarrow \sum_{a\in\mathcal{A}} w^*_a$; $\quad \hat{J} \leftarrow M/\max\{1,S\}$; $\quad d_a \leftarrow w^*_a/\max\{1,S\}$ for $a \in \mathcal{A}$, $d_{\mathbf{0}} \leftarrow 1 - \sum_{a\in\mathcal{A}} d_a$
Return $\hat{J}$ and the certificate d (Proposition 1)

B Proofs

Proof of Proposition 1. The weights w_a, $a \in \mathcal{A}$, are positive, so $d_a \ge 0$ for $a \in \mathcal{A}$. Moreover $\sum_{a\in\mathcal{A}} d_a = S/\max\{1,S\} \le 1$, so $d_{\mathbf{0}} \ge 0$ and $\sum_a d_a = 1$. Since $\mathbf{0}\mathbf{0}^\top = 0$,

$$\sum_a d_a\, aa^\top = \sum_{a\in\mathcal{A}} \frac{w_a}{\max\{1,S\}}\, aa^\top + d_{\mathbf{0}}\, \mathbf{0}\mathbf{0}^\top = \frac{M}{\max\{1,S\}} = \hat{J}. \tag{9}$$

Thus d is a probability distribution over worlds that realizes $\hat{J}$, which is exactly Definition 1. The argument uses neither the value of τ nor the optimality of M. □

Proof of Proposition 2. Let $F(X) = \|J - X\|_F^2$ and $R = J - M$. At termination two facts hold. (a) Every $a \in \mathcal{A}$ has $a^\top R a = 0$, because the final weights minimize the unconstrained least squares objective over $\mathcal{A}$; hence $\langle R, M \rangle = \sum_{a \in \mathcal{A}} w_a \, a^\top R a = 0$. (b) Every $a \notin \mathcal{A}$ has $a^\top R a < \tau$, because (7) is solved to global optimality and $g^\star < \tau$. Together, $a^\top R a < \tau$ for all $a \in \{0,1\}^N$ (for $a = \mathbf{0}$ the value is 0).

Now let $M' = \sum_a w'_a \, aa^\top \in \mathcal{K}$ be arbitrary. Expanding and using (a),

$$F(M') = \|R + (M - M')\|_F^2 = F(M) - 2 \sum_a w'_a \, a^\top R a + \|M - M'\|_F^2. \tag{10}$$

Every nonzero world has at least one entry equal to one, so $\sum_{a \neq \mathbf{0}} w'_a \leq \sum_a w'_a \|a\|_1 = \operatorname{tr}(M')$, and by (a)–(b) $\sum_a w'_a \, a^\top R a \leq \tau \operatorname{tr}(M')$. Therefore

$$F(M) - F(M') + \|M - M'\|_F^2 \leq 2\tau \operatorname{tr}(M'). \tag{11}$$

Taking $M' = M^\star$ and using $F(M^\star) \leq F(M)$ gives both bounds in (8) in terms of $\operatorname{tr}(M^\star)$. For the last inequality, $M^\star$ is the Euclidean projection of J onto the closed convex cone $\mathcal{K}$, so $\langle J - M^\star, M^\star \rangle = 0$ and hence $\|M^\star\|_F \leq \|J\|_F$; by Cauchy–Schwarz, $\operatorname{tr}(M^\star) = \langle I, M^\star \rangle \leq \sqrt{N} \|M^\star\|_F \leq \sqrt{N} \|J\|_F$.

With the stopping test $g^\star \leq 0$, Algorithm 1 coincides with the Lawson–Hanson NNLS algorithm on the (finite) set of 2^N columns, which terminates after finitely many iterations at an optimal solution (Lawson & Hanson, 1995). Finally, if $S \leq 1$ then $\hat{J} = M$, and since $\mathcal{C} \subset \mathcal{K}$ we have $\min_{P \in \mathcal{C}} F(P) \geq F(M^\star) \geq F(M) - 2\tau \operatorname{tr}(M^\star)$. □

C Sensitivity to the Tolerance τ

Table 4 reports the full results summarized in Section 5.4.

Table 4: RMSE and projection runtime (seconds) as a function of the tolerance τ for premises 1–8. "Raw" is the RMSE of the unprojected LLM outputs. Every projected output is exactly coherent regardless of τ.

			RMSE after projection				Runtime (s)			
Premise	Model	Raw RMSE	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
1	Gemma 2	0.3639	0.1875	0.1875	0.1875	0.1875	12.5	14.8	15.0	15.1
	GPT-OSS	0.1101	0.1196	0.1197	0.1197	0.1197	7.5	10.8	14.3	14.3
	GLM-4.5	0.2158	0.1541	0.1542	0.1542	0.1542	26.1	27.3	27.3	27.3
	Qwen 2.5	0.3798	0.1899	0.1900	0.1900	0.1900	6.4	8.6	8.7	8.7
2	Gemma 2	0.3898	0.1584	0.1584	0.1584	0.1584	67.4	72.9	73.8	73.7
	GPT-OSS	0.0988	0.1183	0.1185	0.1185	0.1185	18.0	25.8	28.9	29.1
	GLM-4.5	0.2260	0.1434	0.1434	0.1434	0.1434	36.8	38.4	39.6	39.9
	Qwen 2.5	0.3888	0.1837	0.1837	0.1837	0.1837	56.9	62.4	62.6	63.0
3	Gemma 2	0.3581	0.1868	0.1867	0.1867	0.1867	8.9	10.9	11.4	11.5
	GPT-OSS	0.1147	0.1169	0.1164	0.1164	0.1164	16.1	22.3	25.2	25.1
	GLM-4.5	0.2133	0.1385	0.1385	0.1385	0.1385	22.6	25.3	25.8	25.8
	Qwen 2.5	0.3175	0.1831	0.1833	0.1833	0.1833	8.0	8.6	8.6	8.6
4	Gemma 2	0.3678	0.1663	0.1663	0.1663	0.1663	79.4	84.0	85.4	85.5
	GPT-OSS	0.0691	0.0904	0.0905	0.0905	0.0905	16.4	21.3	23.1	23.9
	GLM-4.5	0.2482	0.1277	0.1277	0.1277	0.1277	40.1	45.2	46.1	46.1
	Qwen 2.5	0.3004	0.1338	0.1339	0.1339	0.1339	47.9	51.1	52.6	52.6
5	Gemma 2	0.3988	0.1856	0.1856	0.1856	0.1856	151.9	175.5	177.2	177.5
	GPT-OSS	0.0735	0.1346	0.1347	0.1347	0.1347	28.6	46.9	53.3	54.7
	GLM-4.5	0.2300	0.1440	0.1440	0.1440	0.1440	111.0	121.0	121.9	121.9
	Qwen 2.5	0.3005	0.1649	0.1649	0.1649	0.1649	96.8	108.4	108.5	108.5
6	Gemma 2	0.3412	0.2428	0.2428	0.2428	0.2428	8.5	8.5	8.5	8.5
	GPT-OSS	0.2404	0.1766	0.1769	0.1769	0.1769	7.4	8.2	8.1	8.2
	GLM-4.5	0.3408	0.2127	0.2126	0.2126	0.2126	2.1	3.1	3.1	3.1
	Qwen 2.5	0.4069	0.3666	0.3666	0.3666	0.3666	7.3	7.6	7.6	7.6
7	Gemma 2	0.4023	0.2717	0.2717	0.2717	0.2717	12.6	16.1	16.5	16.5
	GPT-OSS	0.2255	0.1790	0.1792	0.1792	0.1792	12.6	13.3	14.1	14.1
	GLM-4.5	0.4190	0.2255	0.2255	0.2255	0.2255	21.4	21.6	21.8	21.7
	Qwen 2.5	0.4965	0.2247	0.2247	0.2247	0.2247	9.4	9.9	10.5	10.5
8	Gemma 2	0.3203	0.2110	0.2111	0.2111	0.2111	33.4	36.4	36.6	36.6
	GPT-OSS	0.2999	0.1985	0.1985	0.1985	0.1985	33.8	37.4	39.3	39.1
	GLM-4.5	0.3287	0.1689	0.1689	0.1689	0.1689	24.4	28.3	29.4	29.1
	Qwen 2.5	0.4050	0.2109	0.2108	0.2108	0.2108	24.4	25.8	26.5	26.5

D Qualitative Example

To illustrate what the projection changes, consider premise 2 (a fair coin flipped three times) and the base events E_1: `tails occurs more than two times`; E_2: `tails occurs zero times`; E_3: `heads occurs three times`. The true co-occurrence matrix $J^\star$, the raw predictions J, and the projection $\hat{J}$ restricted to these three events are

$$J^\star = \begin{pmatrix} 0.125 & 0 & 0 \\ 0 & 0.125 & 0.125 \\ 0 & 0.125 & 0.125 \end{pmatrix}, \qquad J = \begin{pmatrix} 0.125 & 0 & 0.25 \\ 0 & 0.125 & 0.125 \\ 0.25 & 0.125 & 0.125 \end{pmatrix}, \qquad \hat{J} = \begin{pmatrix} 0.27 & 0 & 0.14 \\ 0 & 0.27 & 0.05 \\ 0.14 & 0.05 & 0.21 \end{pmatrix}. \tag{12}$$

The raw matrix is infeasible even locally: it assigns the conjunction $E_1 \cap E_3$ a probability of 0.25, larger than both of its marginals (0.125 each), even though the two events are mutually exclusive. The projection lowers this entry to 0.14 and simultaneously adjusts the marginals and the other intersections. Coherence is restored at the cost of moving some entries away from the truth; over all statements, projection still lowers RMSE for most settings (Table 3). These additional changes are necessary rather than incidental: correcting a single violated entry while freezing all others does not in general produce a coherent matrix, because coherence constrains all entries jointly (Section 3).