Bounding Compositional Incoherence in Foundation Models

Anany Kotawala [1]

Abstract

Multi-component foundation-model agents compose several LLM calls into a joint belief; even when each component is internally coherent on its own questions, the assembled belief need not be. We define the *compositional residual* $\varepsilon^\star$, the L_2 distance from the composed quote to the joint coherent polytope, computable at runtime from system output and the explicitly specified cross-component constraints alone. A product-structure dichotomy characterises when local coherence implies global coherence, and a Rayleigh-quotient prediction recovers the typical residual from the specialist panel covariance. Across 1,876 ensemble cliques on four LLMs, $\varepsilon^\star > 0$ on 33–94% of cliques by relation type; a hierarchical Boyle–Dykstra projection eliminates it at $1 \times m$ elicitation cost while retaining specialist routing, improves Brier against resolved labels by up to 0.014 ($p < 10^{-15}$), and yields $+0.115$ nats per bet on 1,770 resolved bets. A planner-discretion harness confirms $\varepsilon^\star > 0$ on $20/20$ partitions under deployed routing.

1. Introduction

Foundation-model evaluation routinely reports per-question accuracy, calibration, and proper-scoring statistics; formal, instance-wise guarantees on *system-level* performance under composition of multiple model calls are comparatively scarce. In multi-component agents, planners route retrieval, arithmetic, and probability assessment to specialist subagents or tools, and each component handles only part of a joint question. Even when every component is calibrated and internally coherent on its assigned questions, the aggregated belief need not be: a research component emitting $P(\text{Republican})=0.6$ and a forecasting component emitting $P(\text{Democrat})=0.6$ produce a 1.2-mass quote that no probability measure can assign, inducing a Dutch-book exposure (de Finetti, 1937) that arises strictly *between* components. Per-component coherence does not, in general, repair the composed system; we call this regime *locally coherent, globally incoherent*.

Per-component calibration, self-consistency (Wang et al., 2023), and conformal prediction (Angelopoulos & Bates, 2021) act on individual model outputs and preserve only per-output coherence properties; cross-component logical constraints are invisible to all three. We identify the *compositional residual* $\varepsilon^\star$, the L_2 distance from the composed quote to the joint coherent polytope, as a runtime, distribution-free certificate of system-level coherence under owner-selected aggregation; it is computable from the composed quote and the cross-component constraints alone. As a constructive repair preserving specialist routing, we use a hierarchical Boyle–Dykstra projection. We pair these with closed-form bounds, an anytime-valid sequential test, and a decomposable benchmark whose hardness is predicted by the theory.

On the theoretical side, we establish a product-structure dichotomy (Theorem 3.3) showing that under owner-selected coordinate aggregation, component coherence guarantees system coherence for all inputs if and only if the joint coherent polytope factorizes as a Cartesian product of local polytopes; under any tighter coupling, locally coherent component forecasts exist whose composition is globally incoherent, and $\varepsilon^\star$ certifies this failure mode (Cor. 3.4). Proposition 3.8 bounds the residual deterministically by inter-component disagreement, and Cor. 3.9 converts the existential dichotomy into a Rayleigh-quotient magnitude prediction keyed to the binding-constraint normal and the diagonal specialist covariance. The same construction yields a closed-form *Brier tax* quantifying the per-clique expected loss attributable to incoherent composition (§3.4). On the operational side, $\varepsilon^\star$ yields a bound $\mathrm{Exposure}^\star \le \sqrt{m^\star}\,\varepsilon^\star$ (Cor. 3.5) on system-level Dutch-book exposure, and admits a constructive hierarchical repair by Boyle–Dykstra cyclic projection (Theorem 3.10). Against any $p^\star \in \mathcal{M}^\star$, it gives a system-level Pythagorean Brier improvement (Cor. 3.6) together with a sharp condition-and-reversal under $\{0,1\}$ resolved labels (Cor. 3.7). The corresponding e-process gives an anytime-valid coherence test with uniform Type-I control at any stopping time (§3.7, Theorem D.2). The L_2-projection, Boyle–Dykstra cyclic-projection, and

[1]Princeton University, Princeton, NJ, USA. Correspondence to: Anany Kotawala <akotawala@princeton.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

Pythagorean-Brier machinery we apply are classical; the contribution is the operational reframing of $\varepsilon^\star$ as a runtime certificate for the joint coherent polytope, the dichotomy of Theorem 3.3 characterising when local coherence suffices, the falsifiable Rayleigh-quotient magnitude prediction of Cor. 3.9, and the decomposable benchmark whose hardness ordering the theory predicts.

Empirically, we evaluate 1,876 ensemble cliques drawn from Paleka (Paleka et al., 2025) and Polymarket across four contemporary foundation models, stratified by four logical-relation classes. The empirical hardness ordering (partition $>$ negation $>$ disjunction $>$ conjunction) tracks how tightly each polytope constrains the joint quote, as predicted by Theorem 3.3, and Cor. 3.9's magnitude prediction matches the observed expected residual within 3–7% on three of the four relation classes (the conjunction under-shoot at $0.83\times$ is itself predicted by the Cor. 3.9 interior-$\bar{\Pi}$ regime). The exposure bound $\sqrt{m^\star}\varepsilon^\star$ averages 0.137 under naive composition (Figure 1b) and is reduced to the QP solver floor by hierarchical projection; we report per-cell Brier with Diebold–Mariano significance, leakage control, a tool-augmented A/B, a same-model decoupling control, and a reproducibility manifest.

2. Preliminaries

Throughout, starred objects denote the composed system, while subscripts denote individual components.

Cliques and the coherent polytope. A clique C is a tuple $(Q_1, \ldots, Q_m, R)$ of m Bernoulli questions and logical relations R. By de Finetti (1937), the set $\mathcal{M}_C = \{r \in [0,1]^m : \exists\, \mu \in \Delta(\{0,1\}^m) \text{ consistent with } R\}$ is a closed convex polytope. We denote the L_2 projection $\Pi_C : [0,1]^m \to \mathcal{M}_C$ and the residual $\varepsilon_C(\hat{p}) := \|\hat{p} - \Pi_C(\hat{p})\|_2$.

K-sample empirical marginal. A forecaster supplies K i.i.d. samples per question, yielding $\hat{p} \in [0,1]^m$ with $\hat{p}_j = K^{-1}\sum_k Y_{j,k}$ and population limit $p_F := \lim_{K\to\infty} \hat{p}$. Generically $\hat{p} \notin \mathcal{M}_C$ even when $p_F \in \mathcal{M}_C$ (due to finite-sample fluctuation).

Within-component JCD. Joint-Coherent Decoding is the L_2 projection of K-sample LLM marginals onto the local coherent polytope $\mathcal{M}_C$. The compositional layer treats JCD as the local repair operator; it inherits a Pythagorean Brier improvement (Theorem A.1, App. A) and a finite-state no-Dutch-book bound via the Fundamental Theorem of Asset Pricing (FTAP; Cor. A.2) bounded by $m \cdot \varepsilon^{\mathrm{KKT}} \leq 1.4\times 10^{-5}$ at OSQP default tolerance.

Multi-component agent. The agent has k sub-models indexed by $a \in [k]$. Sub-model a emits an empirical marginal $\hat{p}^{(a)} \in [0,1]^{m_a}$ on its local question set $\mathcal{Q}_a$ with local polytope $\mathcal{M}_a \subseteq [0,1]^{m_a}$; component-level JCD produces $\Pi_a(\hat{p}^{(a)}) \in \mathcal{M}_a$. The agent's joint question set is $\mathcal{Q}^\star = \bigcup_a \mathcal{Q}_a$ with size $m^\star = |\mathcal{Q}^\star|$ and joint polytope $\mathcal{M}^\star \subseteq [0,1]^{m^\star}$, which respects (i) each local relation lifted to $\mathcal{Q}^\star$ and (ii) all cross-component *coupling constraints* $\mathcal{C}$: identifications of shared questions, cross-component logical relations like $Q^\star = Q_a \wedge Q_b$, and partitions spanning components.

Aggregator. An aggregator $\mathcal{A} : \prod_a [0,1]^{m_a} \to [0,1]^{m^\star}$ assembles the agent's quoted marginal vector. The aggregator $\mathcal{A}$ is the owner-selected coordinate selector of Definition 3.2; this covers the specialist-routing and tool-call composition patterns used in our experiments. We focus on owner-selection because it is the practical pattern under compute budgets: each component owns a subset of joint coordinates and there is no redundant elicitation across specialists. A coord-wise averaging aggregator $\mathcal{A}^{\mathrm{avg}}(\Pi_1, \ldots, \Pi_k) = k^{-1}\sum_a \Pi_a$ on the same inputs would already lie in $\mathcal{M}^\star$ by convexity, so the failure mode of Theorem 3.3 is structurally absent under averaging; we verify this empirically (Section 4). The cost of averaging is $k\times$ elicitation per coordinate, which is the regime specialist-routing is designed to avoid.

The two candidate workflows. With $\hat{p} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})$ and the aggregator $\mathcal{A}$ fixed, two natural projections into $\mathcal{M}^\star$ arise:

$$\text{Local-then-global:} \quad \Pi^\star(\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)})), \tag{1}$$

$$\text{Direct global:} \quad \Pi^\star(\mathcal{A}(\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})). \tag{2}$$

The substantive question is when, and by how much, these differ.

3. Method

Throughout this section we assume the coupling set $\mathcal{C}$ is explicitly specified; the regime of implicit $\mathcal{C}$ recovered from agent transcripts is discussed in §6.

3.1. When local coherence composes

Definition 3.1 (Compositional residual)**.** For inputs $\hat{p} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})$ and aggregator $\mathcal{A}$, the *compositional residual* is

$$\varepsilon^\star(\hat{p}) := \big\|\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)}) - \Pi^\star\big(\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)})\big)\big\|_2.$$

It is the L_2 distance from the locally-repaired composed quote to the joint coherent polytope $\mathcal{M}^\star$. In words, it measures how far the locally repaired, assembled quote remains from global coherence.

Lifted local feasible sets. For each component a, lift $\mathcal{M}_a \subseteq [0,1]^{m_a}$ to $\mathcal{M}_a^{\uparrow} \subseteq [0,1]^{m^\star}$ by leaving the non-component coordinates free. Write $\mathcal{M}^{\boxtimes} := \bigcap_a \mathcal{M}_a^{\uparrow}$ for the joint feasibility implied by the local constraints alone (no cross-component coupling). $\mathcal{M}^\star \subseteq \mathcal{M}^{\boxtimes}$ always; equality holds iff the coupling set $\mathcal{C}$ is trivial.

Definition 3.2 (Owner-selected aggregation)**.** Fix the joint question set $\mathcal{Q}^\star$. An ownership map $\mathrm{owner} : \mathcal{Q}^\star \to [k]$ assigns each joint coordinate to a single component. If multiple components originally quote the same question j, we keep separate local coordinates $p_{a,j}$ and $p_{b,j}$ in the product input space before owner selection, and record the equality $p_{a,j} = p_{b,j}$ in the coupling set $\mathcal{C}$. The aggregator $\mathcal{A}$ selects the coordinate owned by $\mathrm{owner}(j)$. Aggregation only *selects*; coupling *enforces agreement*.

Theorem 3.3 (Product-structure dichotomy)**.** *Under owner-selected coordinate aggregation $\mathcal{A}$, local coherence guarantees global coherence* for all inputs $\hat{p}$ *if and only if $\mathcal{M}^\star = \mathcal{M}^{\boxtimes}$. When $\mathcal{M}^\star = \mathcal{M}^{\boxtimes}$, projection decomposes blockwise:*

$$\Pi^\star(\mathcal{A}(\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})) = \big(\Pi_1 \hat{p}^{(1)}, \ldots, \Pi_k \hat{p}^{(k)}\big), \qquad (3)$$

so $\varepsilon^\star \equiv 0$. When $\mathcal{M}^\star \subsetneq \mathcal{M}^{\boxtimes}$, there exists $r \in \mathcal{M}^{\boxtimes} \setminus \mathcal{M}^\star$, and locally coherent component forecasts realizing r yield a composed quote with $\varepsilon^\star > 0$.

Proof (forward). If $\mathcal{M}^\star = \mathcal{M}^{\boxtimes}$, the squared L_2 norm decomposes: $\|\hat{p} - r\|^2 = \sum_a \|\hat{p}^{(a)} - r^{(a)}\|^2$ for any $r \in \mathcal{M}^{\boxtimes}$, so the joint minimization separates into k independent local minimizations. Hence $\Pi^\star(\hat{p}) = (\Pi_1 \hat{p}^{(1)}, \ldots, \Pi_k \hat{p}^{(k)})$ which equals the aggregator's output on this product space, and $\varepsilon^\star \equiv 0$. This is the block-product structure of L_2 projection in Hilbert space (Bauschke & Combettes, 2017, Prop. 29.3). □

Proof (reverse, existential). Take any $r \in \mathcal{M}^{\boxtimes} \setminus \mathcal{M}^\star$ and lift it to component restrictions; owner-selection then reassembles r from locally coherent inputs, and $r \notin \mathcal{M}^\star$ closed gives $\varepsilon^\star = \|r - \Pi^\star(r)\|_2 > 0$. Full construction in App. B. □

The forward direction is the block-product factorization of L_2 projection in Hilbert space (Bauschke & Combettes, 2017, Prop. 29.3); the contribution is operational, identifying when local coherence is sufficient and when $\varepsilon^\star$ is an appropriate online certificate. We also refer to Theorem 3.3 as the *non-commutation theorem*, since it characterizes exactly when local-then-global L_2 projection commutes with the aggregator (and when it does not). The reverse direction is existential: it shows local coherence is not sufficient under coupling, but does not characterize the typical magnitude of $\varepsilon^\star$. Section 5 estimates this empirically.

Corollary 3.4 (Zero-residual characterization)**.** *Even under coupling, $\varepsilon^\star = 0$ if and only if $\mathcal{A}(\Pi_1 \hat{p}^{(1)}, \ldots, \Pi_k \hat{p}^{(k)}) \in \mathcal{M}^\star$. The compositional certificate is positive on a given input exactly when local repairs do not satisfy the cross-component constraints.*

Proof. $\Pi^\star$ is the identity on $\mathcal{M}^\star$ by definition; $\|x - \Pi^\star(x)\|_2 = 0$ iff $x \in \mathcal{M}^\star$. □

3.2. Exposure interpretation

Corollary 3.5 (System-level FTAP exposure bound)**.** *Treat the coordinates of $\mathcal{A}(\Pi_1, \ldots, \Pi_k)$ as quoted unit-pay prices on the $m^\star$ Bernoulli contracts of the joint clique $C^\star$. If $\varepsilon^\star > 0$, the finite-state FTAP yields a Dutch-book portfolio against the joint quote. Moreover, under the unit-stake Logarithmic Market Scoring Rule (LMSR; Hanson, 2003) exposure statistic used in our experiments,* $\mathrm{Exposure}^\star \leq \sqrt{m^\star}\,\varepsilon^\star$. *Every individual sub-model's component-level Dutch-book exposure is zero by construction up to the QP tolerance (since $\Pi_a(\hat{p}^{(a)}) \in \mathcal{M}_a$); any positive joint exposure is therefore attributable to cross-component incoherence, modulo the numerical solver floor.*

Proof. If the quoted vector lies outside $\mathcal{M}^\star$, the separating hyperplane form of de Finetti's finite-state FTAP gives a portfolio with nonnegative payoff in every coherent state and strictly positive price discrepancy. For the relation classes used in our experiments, the unit-stake exposure statistic is the positive violation of a defining halfspace (or a sum of mutually exclusive positive violations). If $h(x) = (\langle a, x\rangle - b)_+$ is any such violation and $x^\Pi = \Pi^\star(x)$, then $h(x) \leq |\langle a, x - x^\Pi\rangle| \leq \|a\|_2 \|x - x^\Pi\|_2$. For negation, conjunction, disjunction, and partition constraints $\|a\|_2 \leq \sqrt{m^\star}$ after the relation is embedded in the $m^\star$ joint coordinate system, yielding $\mathrm{Exposure}^\star \leq \sqrt{m^\star}\varepsilon^\star$. Each $\Pi_a(\hat{p}^{(a)}) \in \mathcal{M}_a$ has zero local residual to $\mathcal{M}_a$ by definition, so the per-component exposure diagnostic is zero up to numerical solver tolerance. □

Corollary 3.6 (System-level Pythagorean Brier improvement)**.** *Let $p^\star \in \mathcal{M}^\star$ be the true joint marginal vector and let r be any composed quote. Then $\|\Pi^\star(r) - p^\star\|_2^2 \leq \|r - p^\star\|_2^2 - \|r - \Pi^\star(r)\|_2^2$. In particular, taking $r = \mathcal{A}(\Pi_1 \hat{p}^{(1)}, \ldots, \Pi_k \hat{p}^{(k)})$ yields a deterministic, sample-path Brier-improvement guarantee for hierarchical JCD whose slack $\|r - \Pi^\star(r)\|_2^2 = (\varepsilon^\star)^2$ is largest exactly where the compositional residual is largest.*

Proof. The Hilbert-projection argument of Theorem A.1 applied to $\Pi^\star$ on $\mathcal{M}^\star$. □

Corollary 3.7 (Brier lift under $\{0,1\}$ labels: condition and reversal)**.** *Let $y \in \{0,1\}^{m^\star}$ be a resolved label vector and let $p^\star := \mathbb{E}[y]$ denote the true marginal. For any composed*

quote r, $\mathbb{E}_y[\mathrm{Brier}(r,y) - \mathrm{Brier}(\Pi^\star(r),y)] = \|r - p^\star\|_2^2 - \|\Pi^\star(r) - p^\star\|_2^2$. If $p^\star \in \mathcal{M}^\star$, this difference is at least $\|r - \Pi^\star(r)\|_2^2 = (\varepsilon^\star)^2 \geq 0$ by Cor. 3.6; hierarchical JCD improves expected Brier deterministically. If $p^\star \notin \mathcal{M}^\star$ (label incoherence: a fraction of resolutions violate the assumed logical structure), the difference can be negative; projection can amplify the error against $p^\star$.

The condition $p^\star \in \mathcal{M}^\star$ is the precise assumption under which Cor. 3.6 transfers from the geometric to the predictive setting. Empirically (§5), the lift holds with $p < 10^{-15}$ on negation, conjunction, and partition; on disjunction the gain is borderline ($\Delta\mathrm{Brier} = +0.0027$, $p = 0.07$), diagnosing a small fraction of disjunction cliques whose resolved labels deviate from the strict $A \vee B$ relation: this is exactly the $p^\star \notin \mathcal{M}^\star$ regime in which Cor. 3.7 predicts the lift can reverse.

3.3. Disagreement controls the residual

Proposition 3.8 (Disagreement upper bound on $\varepsilon^\star$)**.** *Let $r \in \mathcal{M}^\star$ be any* globally coherent reference forecast*, that is, any joint quote that already lies in the joint polytope. Under owner-selected coordinate aggregation,*

$$\varepsilon^\star(\hat{p}) \;\leq\; \big\|\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)}) - r\big\|_2.$$

A natural choice for r is the JCD-projected forecast $\Pi_\beta(\hat{p}^{(\beta)})$ of any component β that itself lies in $\mathcal{M}^\star$, for instance a component routed the entire joint clique. The right-hand side is then the L_2 disagreement between the composed quote and the reference. It vanishes when all coordinates are assigned to the same component, so inter-component disagreement on disputed coordinates is a measurable proxy that both explains and upper-bounds the residual.

Proof. By the variational characterization of L_2 projection onto a closed convex set, $\varepsilon^\star = \|x - \Pi^\star(x)\|_2 \leq \|x - r\|_2$ for any $r \in \mathcal{M}^\star$, where $x = \mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)})$. □

Let Δ_j denote the per-coordinate disagreement between the assigned LLM and a globally-coherent reference β. For each relation class the bound takes a closed form keyed to the constraint normal: $|\Delta_1 + \Delta_2|/\sqrt{2}$ on negation, $|\sum_j \Delta_j|/\sqrt{m^\star}$ on partition, and $\|\Delta\|_2$ on the Fréchet conjunction/disjunction polytope (full derivations in App. C).

Empirically the bound holds on 100% of the 1,876 cliques. Mean bound-to-residual ratios are 1.30 (negation), 2.34 (disjunction), 3.00 (conjunction), and 4.39 (partition); the bound is tightest where the constraint normal aligns with the disagreement direction.

3.4. Quantitative magnitude prediction

Under uniform random owner-selection, the expected residual admits a closed form keyed to the empirical specialist covariance restricted to the constraint normal that defines the joint polytope. Let $\bar{\Pi} := k^{-1}\sum_a \Pi_a \in \mathcal{M}^\star$ (by convexity) and $D := \mathrm{diag}(\Sigma_\Pi)$ where $\Sigma_\Pi := k^{-1}\sum_a (\Pi_a - \bar{\Pi})(\Pi_a - \bar{\Pi})^\top$. Independent assignment across coordinates makes D the effective covariance: $\mathbb{E}_\sigma[(x - \bar{\Pi})(x - \bar{\Pi})^\top] = D$.

Corollary 3.9 (Magnitude of $\varepsilon^\star$, Rayleigh-quotient form)**.** *Suppose the joint polytope is locally defined by a single binding halfspace with normal a_R at the projection. Under uniform i.i.d. owner-selection,*

$$\mathbb{E}_\sigma\big[(\varepsilon^\star)^2\big] \;=\; \kappa_R \cdot \frac{a_R^\top D\, a_R}{\|a_R\|_2^2}, \tag{4}$$

where $\kappa_R = 1$ when the binding constraint is an equality ($a_R^\top \bar{\Pi} = b_R$). When the binding constraint is an inequality with $\bar{\Pi}$ on the boundary, the law of $a_R^\top(x - \bar{\Pi})$ is the empirical distribution of $\{a_R^\top(\Pi_a - \bar{\Pi})\}_{a\in[k]}$; if this scalar is symmetric around 0 then $\kappa_R = \frac{1}{2}$ exactly, and otherwise $\kappa_R \in [0,1]$ with $\kappa_R \approx \frac{1}{2}$ when the panel is approximately balanced about the boundary. The upper bound $\mathbb{E}_\sigma[(\varepsilon^\star)^2] \leq a_R^\top D a_R/\|a_R\|_2^2$ holds in all inequality cases, and the generic $\mathbb{E}_\sigma[(\varepsilon^\star)^2] \leq \mathrm{tr}(D)$ holds without further assumptions. For the relation classes of Section 4: $a_{\mathrm{NEG}} = (1,1)$ (equality, $\kappa = 1$); $a_{\mathrm{PARTITION}} = \mathbf{1}_{m^\star}$ (equality, $\kappa = 1$); $a_{\mathrm{AND}}, a_{\mathrm{OR}}$ are Fréchet halfspace normals (inequality, $\kappa \approx \frac{1}{2}$ under the symmetric-panel approximation; conjunction's empirical $\kappa \approx 0.83 \cdot \frac{1}{2}$ reflects $\bar{\Pi}$ off the boundary, see §5).

Proof sketch. Equality: $\varepsilon^\star = |a_R^\top(x - \bar{\Pi})|/\|a_R\|$, square and use $\mathbb{E}_\sigma[a_R^\top(x - \bar{\Pi})(x - \bar{\Pi})^\top a_R] = a_R^\top D a_R$. Inequality with $\bar{\Pi}$ on the boundary: $\varepsilon^\star = (a_R^\top(x - \bar{\Pi}))_+/\|a_R\|$; for a zero-mean random variable X, $\mathbb{E}[X_+^2] = \frac{1}{2}\mathbb{E}[X^2]$ when X is symmetric, otherwise $\mathbb{E}[X_+^2]/\mathbb{E}[X^2] \in [0,1]$ depending on the upper-tail mass. Full proof and the generic-bound case in App. H.

Combined with Cor. 3.7, this yields a closed-form *Brier tax* $\mathbb{E}[\mathrm{Brier}(r,y) - \mathrm{Brier}(\Pi^\star(r),y)] \geq \kappa_R\, a_R^\top D a_R/\|a_R\|^2$ when $p^\star \in \mathcal{M}^\star$, computable from the panel covariance alone. The three couplings recurring in our experiments are *shared-question disagreement* (two specialists quote the same Q at different prices), *negation across components* ($p_a + p_b = 1$), and *partition across components* ($\sum_a p_a = 1$).

3.5. Hierarchical repair

The repair is constructive. Let $\Pi_\mathcal{C} : [0,1]^{m^\star} \to \mathcal{C}$ denote the L_2 projection onto the coupling set, an intersection of finitely many half-spaces.

Theorem 3.10 (Hierarchical projection convergence)**.** *Let $\{\Pi_1, \ldots, \Pi_k, \Pi_\mathcal{C}\}$ be the family of L_2 projections onto the*

local polytopes $\{\mathcal{M}_a\}$ (lifted to $[0,1]^{m^\star}$) and the coupling set $\mathcal{C}$. Their intersection is the nonempty joint polytope $\mathcal{M}^\star$. The Boyle-Dykstra cyclic-projection iteration (Boyle & Dykstra, 1986) initialized at any $r_0 \in [0,1]^{m^\star}$ generates a sequence converging in L_2 to $\Pi^\star(r_0)$.

Sketch. $\mathcal{M}^\star = \left(\bigcap_a \mathcal{M}_a^\uparrow\right) \cap \mathcal{C}$ where $\mathcal{M}_a^\uparrow$ is $\mathcal{M}_a$ lifted to $[0,1]^{m^\star}$ (free on non-component coordinates). Each $\Pi_\bullet$ is the L_2 projection onto a closed convex set, and each $\mathcal{M}_a^\uparrow$ and $\mathcal{C}$ is closed convex. By Bauschke & Combettes (2017, Thm. 9.31) (Boyle-Dykstra), cyclic L_2 projection over a finite family of closed convex sets converges to the projection onto their intersection. □

Per-iteration cost. Each Π_a and $\Pi_\mathcal{C}$ is a small QP. The Paleka (neg/and/or) and Polymarket (partition) relation classes admit closed-form projections (App. C); we use direct $\Pi^\star$ via OSQP (Stellato et al., 2020) for uniformity across relation classes, since the partition simplex projection and the negation closed form are mathematically equivalent to OSQP's converged solution at the reported tolerance.

3.6. Runtime use

The residual supports three deployment modes that share the same computation and differ only in how the agent acts on it. In *monitor* mode the agent logs $\varepsilon^\star$ and surfaces large residuals to an operator. In *repair* mode it replaces the naive composed quote with $\Pi^\star(\cdot)$ before passing it downstream. By Theorem 3.10 and Cor. 3.5, the repaired quote lies in $\mathcal{M}^\star$ up to solver tolerance and the exposure diagnostic is driven to numerical zero. In *abstain-or-escalate* mode it refuses to act when $\varepsilon^\star > \tau$, a risk–coverage knob whose calibration to budgeted exposure is outside our scope.

3.7. Sequential monitoring: anytime-valid coherence test

In long-horizon agent deployments the agent generates a compositional residual stream $(\varepsilon_t^\star)_{t\geq 1}$ across successive composition steps. The spatial guarantee of Section 3 certifies incoherence at each step; the following temporal guarantee lets an operator certify *persistent* incoherence across the stream without fixing a stopping horizon in advance.

Under the null H_0 that all component forecasters are coherent, the e-process $E_t(\lambda) := \prod_{s=1}^t \exp\big(\lambda(\varepsilon_s^{\star 2} - m_s^\star/(4K_s)) - \lambda^2 m_s^\star/(2K_s)\big)$ is a non-negative $\mathcal{F}_t$-supermartingale (Lemma D.1, App. D). By Ville's inequality (Ville, 1939; Howard et al., 2021), the stopping rule $\tau_\alpha := \inf\{t : E_t(\lambda) \geq 1/\alpha\}$ controls Type-I error at level α uniformly over all stopping times (Theorem D.2). An operator monitors E_t online and escalates the agent pipeline when it crosses $1/\alpha$; no fixed-horizon commitment is required. The full derivation and empirical demonstration on four LLM residual streams appear in Section D.

4. Experimental setup

We evaluate on Paleka (Paleka et al., 2025) (negation, conjunction, disjunction, paraphrase; 134 cliques per checker) and Polymarket (67 partition cliques after leakage control, expanded to 268 ensemble instances by four seeds). Both benchmarks were chosen because they target the regime the theory predicts will be tightest: Paleka was constructed by Paleka et al. (2025) for cross-question consistency checking, and Polymarket's partition events are the unit-mass constraint regime ($\sum_i p_i = 1$) where the polytope coupling is most restrictive. **Models.** Anthropic `claude-haiku-4-5-20251001` (via the Anthropic Messages API), OpenAI `gpt-5.4-mini` and `gpt-5.4-nano` (via Azure OpenAI), and Meta `llama-3.3-70b-versatile` (via Groq). Frontier-panel evaluations additionally use OpenAI `gpt-5.5` (via Azure OpenAI). These are the public API model identifiers as returned by each provider on 30 April 2026; raw responses, prompts, and per-clique residuals are released with the supplementary material. Each specialist produces $K{=}8$ verbalized probability samples per question at temperature 0.7. **Leakage control.** Polymarket events were filtered so that every market in every retained event resolved strictly after our model snapshot date of 30 April 2026; the inclusion cutoff was 1 May 2026, and all retained events resolved on or after that date. Paleka cliques inherit the resolution-status verification of Paleka et al. (2025). Full filtering protocol in Appendix E. **Compositional ensemble:** for each clique with m questions, we draw 4 random seeds; each seed assigns each of the m questions to one of the 4 LLMs uniformly i.i.d.; the agent's quoted marginal on coordinate j is the JCD-projected marginal of the assigned LLM on that coordinate. The source forecast from each LLM is JCD-coherent on the full clique before coordinate selection; the ensemble residual $\varepsilon^\star$ therefore reflects *exclusively* cross-component disagreement.

Benchmark properties. The four relation classes admit closed-form local projections (App. C) and span a range of polytope geometries from a single equality (negation), to a high-dimensional simplex (partition), to the Fréchet box (and/or). The coupling set $\mathcal{C}$ controls the gap between $\mathcal{M}^\star$ and $\mathcal{M}^\boxtimes$, and random vs. structured routing brackets cross-model mixing. Theorem 3.3's prediction $\varepsilon^\star \equiv 0$ when $\mathcal{M}^\star = \mathcal{M}^\boxtimes$ is verified at the QP solver floor by the no-composition reference (Figure 1b, gray dashed), a falsifiable test of the dichotomy.

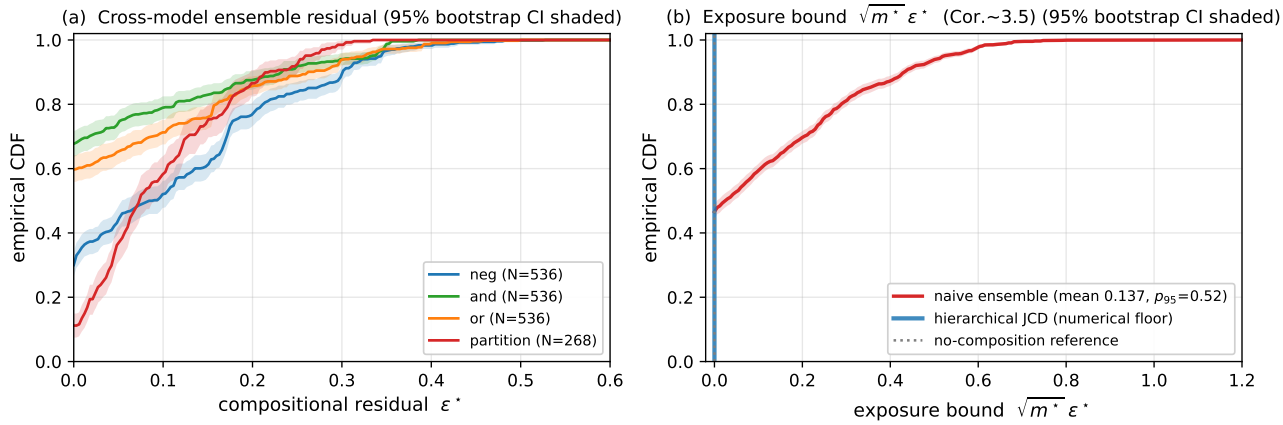

Figure 1. **Compositional residuals and the exposure bound.** *(a)* ECDF of $\varepsilon^\star$ across N=1,876 ensemble cliques, stratified by relation. Each component is per-call JCD-projected, so any positive $\varepsilon^\star$ is cross-component incoherence introduced by composition. *(b)* ECDF of the exposure bound $\sqrt{m^\star}\varepsilon^\star$ (Cor. 3.5) under three regimes: naive composition (mean 0.137), hierarchical JCD (at the OSQP floor), and a single-LLM no-composition reference. Shaded bands are 95% bootstrap CIs (B=1000).

5. Results

The random-assignment experiment ($N=536$/relation $\times 4$ seeds for Paleka, $268=67\times4$ for Polymarket; total 1,876) and structured-routing variant ($N=469$, one deterministic seed) bracket cross-model mixing.

5.1. Compositional residual and controls

Compositional residual under random routing. The compositional residual is strictly positive on 94% (partition), 66% (negation), 43% (disjunction), 33% (conjunction) of cliques (Figure 1a). Mean $\varepsilon^\star$ ranges 0.058 (and) to 0.118 (partition); *none* is visible per-component, where each $\Pi_a \in \mathcal{M}_a$ has zero residual. Hierarchical JCD reduces exposure to 1.7×10^{-14} (OSQP solver floor), matching the no-composition reference (a single LLM on every coordinate; Figure 1b, gray) while retaining specialist routing.

Empirical validation of the magnitude prediction. Cor. 3.9's Rayleigh-quotient prediction $\kappa_R\, a_R^\top D a_R / \|a_R\|^2$ matches the observed expected residual $\mathbb{E}_\sigma[(\varepsilon^\star)^2]$ across all four relation classes (Table 1; per-clique Pearson 0.69–0.89). The conjunction under-shoot reflects $\bar{\Pi}$ typically interior to the binding Fréchet halfspace; on the remaining three relations the prediction is tight to within 3–7%. Both $\varepsilon^\star$ and the $\sqrt{m^\star}\varepsilon^\star$ exposure bound are therefore predictable from the panel covariance alone, before any routing.

Same-model decoupling control. To separate coordinate isolation from cross-model capability divergence, we reran with four fresh-seed runs of one model (Claude-Haiku-4.5, N=80 cliques per relation, 40 partitions). Mean $\varepsilon^\star$ falls to 0.025/0.019/0.040/0.058 on neg/and/or/partition: 22–60% of the residual persists at same-model, with cross-model heterogeneity amplifying by 1.7–4.5×. A greedy-decoding (T=0) control retains $\varepsilon^\star$ = 0.025–0.096 across relations (App. J); on negation and disjunction T=0 amplifies the residual relative to T=0.7, ruling out sampling noise as the source.

Relation	κ_R	Pred./Obs.	Match
NEGATION	1	1.054	95%
PARTITION	1	1.069	94%
DISJUNCTION	$\frac{1}{2}$	1.026	97%
CONJUNCTION	$\frac{1}{2}$	0.830	83%

Table 1. Theory vs. observed expected squared residual. *Pred./Obs.* is the ratio of Cor. 3.9's Rayleigh-quotient prediction to measured $\mathbb{E}_\sigma[(\varepsilon^\star)^2]$ across the 1,876 ensemble cliques; *Match* is $1-|1-$ Pred./Obs.$|$. The conjunction under-shoot is the $\bar{\Pi}$-interior regime predicted by Cor. 3.9.

Compositional Brier against resolved labels. We adopt the convention $\Delta\text{Brier} = \text{Brier}_{\text{JCD}} - \text{Brier}_{\text{naive}}$ throughout, so a negative Δ indicates JCD improvement. Paired over the 1,876 cliques, hierarchical JCD reduces compositional Brier on NEG (ΔBrier$=-0.0137$, $p<10^{-43}$), AND (ΔBrier$=-0.0076$, $p<10^{-16}$), and PARTITION (ΔBrier$=-0.0048$, $p<10^{-23}$). OR is marginal (ΔBrier$=+0.0027$, p=0.07): this is the predicted reversal regime of Cor. 3.7, where label noise on a small fraction of disjunction resolutions pushes $p^\star$ outside $\mathcal{M}^\star$, and Cor. 3.7's condition-and-reversal allows projection to amplify error there rather than reduce it. The OR direction is therefore positive evidence for the theorem's sharpness, not a failure of JCD on disjunction.

Disagreement-driven mechanism. Per-clique $\varepsilon^\star$ regressed on the assignment-aware reference disagreement

$\min_\beta \|x - \Pi_\beta\|_2$ (Prop. 3.8) gives R^2=0.93 on negation (slope 0.83) and R^2=0.58–0.64 on and/or/partition (slope 0.59–0.63). The constraint normal aligns most closely with the disagreement direction on negation; on higher-dimensional polytopes the projection contracts components of disagreement orthogonal to the binding halfspace, loosening the bound.

Robustness to routing topology and retrieval grounding. Under structured routing (one model per relation family; App. L), $\varepsilon^\star > 0$ on 98.5% of negation, 36% partition, 16% disjunction, 12% conjunction cliques; mean $\sqrt{m^\star}\varepsilon^\star = 0.117$ vs. 0.137 under random routing. Retrieval grounding (top-5 DuckDuckGo snippets per specialist) on 30 partitions shifts mean $\varepsilon^\star$ from 0.260 to 0.283; 20/30 partitions are equal-or-worse with retrieval (App. P).

Routing-protocol ladder. On 100 multi-candidate partitions (App. O) we walk the routing protocol from fixed naive owner-selection to autonomous tool-using planners. Prompt-engineering under owner-selection drops mean $\varepsilon^\star$ from 0.235 (ISOLATED) to 0.090 (LISTED) to 0.081 (FULL); ISOLATED → LISTED is significant ($p < 10^{-11}$), LISTED → FULL is not ($p = 0.18$), and 99% of partitions retain $\varepsilon^\star > 10^{-3}$ under FULL. Tool-using planners with delegate-and-submit tools further reduce $\varepsilon^\star$: a non-reasoning planner reaches $\varepsilon^\star = 0.009/0.004$ at $\sim 4.8\times m$ calls; a frontier reasoning planner reaches 0/200 positive residuals at $\sim 6.7\times m$ calls. The hierarchical projection reaches $\varepsilon^\star \to 0$ at $1\times m$ specialist calls (Theorem 3.10).

5.2. Planner-discretion harness

To check that the failure is not an artefact of random assignment, we re-run the experiment with a planner LLM (Claude-Haiku-4.5) that, for each of 20 live-style partitions, picks a specialist per outcome from the four-LLM roster and emits its own per-specialist context block; specialists then sample K=8 verbalized probabilities under the planner's prompt and we compute $\varepsilon^\star$ as in §4. Mean $\varepsilon^\star = 0.113$ versus 0.231 under random assignment on the same partitions (15/20 improve), but $\varepsilon^\star > 0$ on 20/20. The planner's routing histogram (GPT-5.4-mini 50%, Claude-Haiku 31%, Llama-70b 13%, GPT-nano 6%) shows no self-routing.

5.3. Downstream-decision regret

For each of 1,770 resolved ensemble bets from §4 (we drop 106/1,876 where the assigned specialist lacks a forecast for the assigned coordinate, a pipeline artefact on partition events) we convert the composed quote $\hat{p}^\star$ into a proportional allocation $w_i = \max(\hat{p}_i^\star, 0)/\sum_j \max(\hat{p}_j^\star, 0)$ and measure (i) Brier loss against the binary resolution vector and (ii) the realised log-payoff $\log w_{\text{winner}}$ on the 1,250 bets with a unique YES. The proportional rule rescales away partition-mass excess and is favourable to naive composition, so the JCD advantage reported below lower-bounds what stricter rules (e.g. truncated Kelly) would show.

Hierarchical JCD lowers mean Brier ($\Delta\text{Brier} = -0.018$, paired-bootstrap 95% CI over (clique, seed): $[-0.021, -0.015]$) and improves mean log-payoff by $+0.115$ nats per bet ($[+0.056, +0.187]$). JCD matches the single-LLM oracle's gain ($\Delta\text{Brier} = -0.017$, CI $[-0.035, -0.000]$) with a confidence interval $\sim 8\times$ narrower. The effect concentrates on partitions (naive loses 0.60 nats/bet vs. JCD, CI $[+0.15, +1.15]$, n=162). Brier is allocation-rule independent; log-payoff is rule-dependent and shrinks to $+0.006$ nats under bettors that themselves coherentise the quote (truncated Kelly, max-entropy). The Pythagorean Brier improvement of Cor. 3.6 is thus realised on resolved labels, not only on the synthetic $\varepsilon^{\star 2}$ slack.

5.4. Per-component layer (summary)

The local JCD layer behaves as Theorem A.1 predicts. Sorting Paleka cliques by pre-projection ε, JCD's Brier gain stratifies to the most-incoherent quartile (0.040 at m=2, 0.013 at m=3; 0.12 on Polymarket at $m \geq 4$); other quartiles show no detectable gain (Figure 3). Stacked with per-question Platt calibration (Platt, 1999) (baseline B6 of App. G), JCD reduces mean Brier on every (model, checker) cell, by 0.037 on Paleka and 0.046 on Polymarket; Diebold–Mariano $p < 10^{-4}$ on every model. A Murphy decomposition (Table 7) attributes the gain to calibration (36% reliability reduction) rather than over-extremization. Per-relation B6 gains on Paleka (-21% and, -37% or, -4% neg, -1% paraphrase) track local-polytope restrictiveness.

Reproducibility. Code, prompts, sample dumps, per-clique JCD residuals, the ensemble-assignment seed, and the full hierarchical-projection pipeline are documented in Section E and released with the supplementary material; K-sweep and the sequential test appear in Appendix D.

6. Conclusion

As foundation-model agents move from per-question benchmarks into multi-step decision pipelines that route subqueries across specialists and tools, the unit of evaluation shifts from the individual model output to the assembled belief on which the agent acts. Per-component calibration, self-consistency, and conformal prediction are well-developed for the former; the latter has lacked an instance-wise, distribution-free guarantee. The compositional residual $\varepsilon^\star$ supplies one. It is computable at deployment time from system output and the cross-component constraints alone, certifies a Dutch-book exposure that no per-component repair can remove, and admits a deterministic projection-based fix.

Concretely, the product-structure dichotomy (Theorem 3.3) characterises when local coherence implies global coherence, and Cor. 3.9 predicts the typical residual from the panel covariance, matching observed $\mathbb{E}_\sigma[(\varepsilon^\star)^2]$ within 3–7% on three of four relation classes and within the predicted interior-$\bar{\Pi}$ regime on conjunction. A hierarchical Boyle–Dykstra projection eliminates $\varepsilon^\star$ at $1 \times m$ elicitation cost; the planner-discretion harness confirms the failure on $20/20$ deployed-routing partitions; and on resolved labels the certificate yields $+0.115$ nats per bet over naive composition. These guarantees are orthogonal to per-output conformal prediction (Angelopoulos & Bates, 2021) and complementary to opinion pooling (Genest & Zidek, 1986; Abbas, 2009), which requires every expert to forecast every coordinate.

Limitations. The Rayleigh-quotient prediction of Cor. 3.9 is exact for negation and partition cliques but conservative when the panel mean $\bar{\Pi}$ lies in the interior of the binding Fréchet halfspace, which produces the $0.83\times$ under-shoot we observe on conjunction.

The principal scope condition is that the coupling set $\mathcal{C}$ is specified explicitly: every theorem in §3 treats $\mathcal{C}$ as a finite, externally given list of half-spaces and equalities over joint coordinates (shared-question identifications, cross-component logical relations, partition-mass constraints). This covers structured tool-use and function-calling deployments, where the planner emits declared output schemas with typed sub-question fields and $\mathcal{C}$ is recoverable from the schema without language understanding. Our planner-discretion harness (§4) and routing-protocol ladder (App. O) target this regime.

It does not cover deployments in which $\mathcal{C}$ is implicit in free-form chain-of-thought or in unstructured planner–tool transcripts. There, $\varepsilon^\star$ remains well-defined and the Boyle–Dykstra repair still applies on any subsequently specified polytope, but the certificate cannot be computed without first recovering $\mathcal{C}$ from the trace. Partial routes to that recovery exist as components: an NLI-style classifier over sub-question pairs, schema-enforcing planner protocols, and symbolic trace logging at the tool-call boundary. A unified inference step remains open; we list it as the first future-work item below.

Finally, our resolved-label experiment relies on Polymarket resolutions, which are noisy on a small fraction of disjunction cliques; this pushes $p^\star$ outside $\mathcal{M}^\star$ on those cliques and is the regime where Cor. 3.7 does not guarantee improvement.

Future work. Three directions follow naturally. First, the coupling set $\mathcal{C}$ is currently provided externally; recovering $\mathcal{C}$ from unstructured planner–tool traces would extend $\varepsilon^\star$ to free-form agent transcripts, where the cross-component constraints are implicit in the prompt structure rather than typed. Second, the L_2/Brier pairing is one instance of a broader matched-Bregman recipe; substituting log-loss or spherical scoring rules yields analogous projections under their respective Bregman geometries (Banerjee et al., 2005), with the residual generalising to a Bregman gap. Third, the spatial certificate of Theorem 3.3 and the temporal e-process of §3.7 compose trivially across composition steps but ignore long-horizon dependence between steps; a sequential certificate that exploits inter-step structure is open.

Impact Statement

A runtime certificate for system-level coherence can be wired into deployed multi-component pipelines as a guardrail, surfacing incoherent compositions before they reach a downstream decision policy. This is a reliability and auditability gain for high-stakes settings such as forecasting, decision support, and regulated probabilistic reasoning. The corresponding risk is over-trust: a small $\varepsilon^\star$ certifies coherence of the quoted belief but says nothing about whether the underlying specialists are calibrated against the true distribution, and a projected quote may be confidently wrong on the very inputs where its components disagree most. Operators deploying the repair should therefore report $\varepsilon^\star$ alongside conventional calibration and coverage diagnostics, not as a substitute for them.

References

Abbas, A. E. A Kullback–Leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37, 2009.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

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.

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.

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.

Genest, C. and Zidek, J. V. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

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

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

McDiarmid, C. On the method of bounded differences. In Siemons, J. (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

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.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

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.

Ville, J. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

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.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A. Within-component projection facts

Theorem A.1 (Pythagorean Brier improvement)**.** *Let $p^* \in \mathcal{M}_C$ be the ground-truth marginal for clique C and $\Pi(\hat{p})$ the L_2 projection of $\hat{p}$ onto $\mathcal{M}_C$. Then for any $\hat{p} \in \mathbb{R}^m$, $\|\Pi(\hat{p}) - p^*\|_2^2 \le \|\hat{p} - p^*\|_2^2 - \|\hat{p} - \Pi(\hat{p})\|_2^2$. The slack $\|\hat{p} - \Pi(\hat{p})\|_2^2$ is the squared distance from coherence; Brier improvement is largest where pre-projection incoherence is largest.*

Proof. Hilbert projection theorem (Bauschke & Combettes, 2017, Thm. 3.16) gives $\langle \hat{p} - \Pi(\hat{p}),\, p^* - \Pi(\hat{p}) \rangle \le 0$. Expand $\|\hat{p} - p^*\|^2$ around $\Pi(\hat{p})$ and drop the non-negative cross term. □

Corollary A.2 (Per-component finite-state FTAP)**.** *Treating $\Pi(\hat{p}) \in \mathcal{M}_C$ as quoted unit-pay prices on the m Bernoulli contracts of C, the finite-state Fundamental Theorem of Asset Pricing rules out a state-independent Dutch book against the ideal projection. Under a numerical QP solution the unit-stake exposure diagnostic is bounded by $m \cdot \varepsilon^{\mathrm{KKT}} \le 1.4 \times 10^{-5}$ at* OSQP *default tolerance.*

B. Proof of Theorem 3.3 (reverse direction)

Suppose $\mathcal{M}^\star \subsetneq \mathcal{M}^\boxtimes$. Pick any $r \in \mathcal{M}^\boxtimes \setminus \mathcal{M}^\star$. For each component a, set $\hat{p}^{(a)}$ to the restriction of r to component a's coordinates: $\hat{p}_i^{(a)} := r_{j(a,i)}$ where $j(a,i)$ is the joint index that component a's i-th local coordinate maps to under Definition 3.2; under shared-question coupling, set $\hat{p}_{i_a}^{(a)} = \hat{p}_{i_b}^{(b)} = r_j$ whenever components a and b both own the duplicate of joint coordinate j. Then each restriction lies in $\mathcal{M}_a$ since $r \in \mathcal{M}^\boxtimes$, so $\Pi_a(\hat{p}^{(a)})$ equals that restriction and the owner-selected aggregator reassembles $\mathcal{A}(\Pi_1, \ldots, \Pi_k) = r$. Since $r \notin \mathcal{M}^\star$ and $\mathcal{M}^\star$ is closed, $\Pi^\star(r) \ne r$ and $\varepsilon^\star = \|r - \Pi^\star(r)\|_2 > 0$.

C. Closed-form local projections

Negation (m=2, $r_1 + r_2 = 1$). $\Pi_{\mathcal{M}_C}(\hat{p}) = \big(\frac{1}{2}(1 + \hat{p}_1 - \hat{p}_2),\ \frac{1}{2}(1 - \hat{p}_1 + \hat{p}_2)\big)$; deterministic Dutch-book on raw $\hat{p}$ is $|\hat{p}_1 + \hat{p}_2 - 1|$.

Conjunction (m=3, r_3=$r_1 \wedge r_2$). $\mathcal{M}_C = \{r \in [0,1]^3 : \max(0, r_1 + r_2 - 1) \le r_3 \le \min(r_1, r_2)\}$ is the Fréchet polytope; $\Pi_{\mathcal{M}_C}$ is the unique fixed point of joint clipping on r_3 and feasibility-preserving adjustments on (r_1, r_2), computed by Dykstra's cyclic projection algorithm (Boyle & Dykstra, 1986) on the six halfspaces (four box, two Fréchet) and converging to machine precision in < 200 iterations. Deterministic Dutch-book on raw $\hat{p}$ is $\max(0, \hat{p}_3 - \min(\hat{p}_1, \hat{p}_2)) + \max(0, \max(0, \hat{p}_1 + \hat{p}_2 - 1) - \hat{p}_3)$.

Disjunction (m=3, r_3=$r_1 \vee r_2$). Bounds $\max(\hat{p}_1, \hat{p}_2) \le r_3 \le \min(1, \hat{p}_1 + \hat{p}_2)$ via $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$; same Dykstra recipe.

Partition ($\sum_i r_i = 1$, $r_i \ge 0$). $\Pi_{\mathcal{M}_C}$ is the simplex projection of Wang & Carreira-Perpiñán (2013) in $O(n \log n)$. Deterministic Dutch-book on raw $\hat{p}$ is $|\sum_i \hat{p}_i - 1|$.

D. Sequential test on the residual stream

In long-horizon agent deployments, the compositional residual stream $(\varepsilon_t^\star)_{t \ge 1}$ can be monitored over time; the per-component test below extends to the system level once a bounded-difference envelope for the chosen aggregation rule is specified. Under coordinate-selection aggregation with disjoint owners, $\varepsilon_t^{\star 2}$ inherits the envelope below with m_t replaced by $m_t^\star$ and K_t replaced by $\min_a K_{a,t}$. The compositional theory of Section 3 is spatial: it controls coherence at a single time step across k components. The residual stream $(\varepsilon_t)_{t \ge 1}$ that an agent generates over time admits a complementary *temporal* guarantee.

Setup. A clique stream $(C_t)_{t \ge 1}$ with sizes m_t and K_t samples per question yields residuals $\varepsilon_t = \|\hat{p}_t - \Pi(\hat{p}_t)\|_2$. Assume the $K_t m_t$ Bernoulli samples at time t are conditionally independent given $\mathcal{F}_{t-1}$, with conditional mean vector $p_F^{(t)}$. Define the null $H_0 :\ p_F^{(t)} \in \mathcal{M}_C^{(t)}$ for all t.

Lemma D.1 (Bounded-difference MGF bound)**.** *Under H_0, $\varepsilon_t^2 \in [0, m_t]$ with conditional mean $\mathbb{E}[\varepsilon_t^2 \mid \mathcal{F}_{t-1}] \le m_t/(4K_t)$, and for every $\lambda > 0$, $\mathbb{E}\big[\exp\big(\lambda(\varepsilon_t^2 - m_t/(4K_t))\big) \mid \mathcal{F}_{t-1}\big] \le \exp\big(\lambda^2 m_t/(2K_t)\big)$.*

Proof. $\varepsilon_t \le \|\hat{p}_t - p_F^{(t)}\|_2$ under H_0 by projection optimality on $\mathcal{M}_C$. The squared norm $Z_t = \|\hat{p}_t - p_F^{(t)}\|_2^2$ is a function of $K_t m_t$ i.i.d. Bernoulli samples with bounded differences $2/K_t$ per sample; McDiarmid's bounded-difference inequality (McDiarmid, 1989) yields the sub-Gaussian MGF bound $\mathbb{E}[\exp(\lambda(Z_t - \mathbb{E}Z_t))] \le \exp(\lambda^2 m_t/(2K_t))$, and $\mathbb{E}Z_t \le m_t/(4K_t)$ from the per-component variance bound. □

Theorem D.2 (Anytime-valid coherence test)**.** *For any $\lambda > 0$, $E_t(\lambda) := \prod_{s=1}^{t} \exp(\lambda(\varepsilon_s^2 - m_s/(4K_s)) - \lambda^2 m_s/(2K_s))$, $E_0 = 1$, is a non-negative $\mathcal{F}_t$-supermartingale under H_0. Hence by Ville's inequality (Ville, 1939; Howard et al., 2021), for every $\alpha \in (0,1)$ and every stopping time τ, $\mathbb{P}\big(\sup_{t \le \tau} E_t(\lambda) \ge 1/\alpha \mid H_0\big) \le \alpha$. The decision rule $\tau_\alpha := \inf\{t : E_t \ge 1/\alpha\}$ controls Type-I error at level α uniformly over stopping rules.*

Proof. Each factor of $E_t(\lambda)$ is non-negative. By Lemma D.1, $\mathbb{E}[E_t/E_{t-1} \mid \mathcal{F}_{t-1}] \le \exp(\lambda^2 m_t/(2K_t)) \exp(-\lambda^2 m_t/(2K_t)) = 1$, the supermartingale property. Ville's inequality yields the maximal bound, uniform over stopping times by optional stopping. □

Tuning λ. For the alternative $\mathbb{E}[\varepsilon_t^2] \to m_t/(4K_t) + \delta$, the bounded-difference-optimal bet is $\lambda^* = \delta K_t/m_t$; we mix over a log grid $\lambda \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$ in our experiments (Figure 2), retaining anytime-valid coverage by linearity (Ramdas et al., 2023).

Empirical demonstration. Figure 2 runs the test on each model's per-clique residual stream (N=603 cliques per model, fixed-seed shuffle). Three of four models reject H_0 at $\alpha \le 10^{-4}$ within $\le$422 cliques (Llama-3.3-70b at $t = 16$, GPT-5.4-nano at $t = 177$, GPT-5.4-mini at $t = 422$); Claude-Haiku-4.5 has mean residual below the H_0-envelope at K=8 and is not rejected within 603 cliques, demonstrating that the test calibrates power against the bounded-difference envelope. The same construction applies to the compositional residual stream $(\varepsilon_t^\star)$ from Section 3 after replacing the per-component envelope by the corresponding envelope for the agent's aggregation rule.

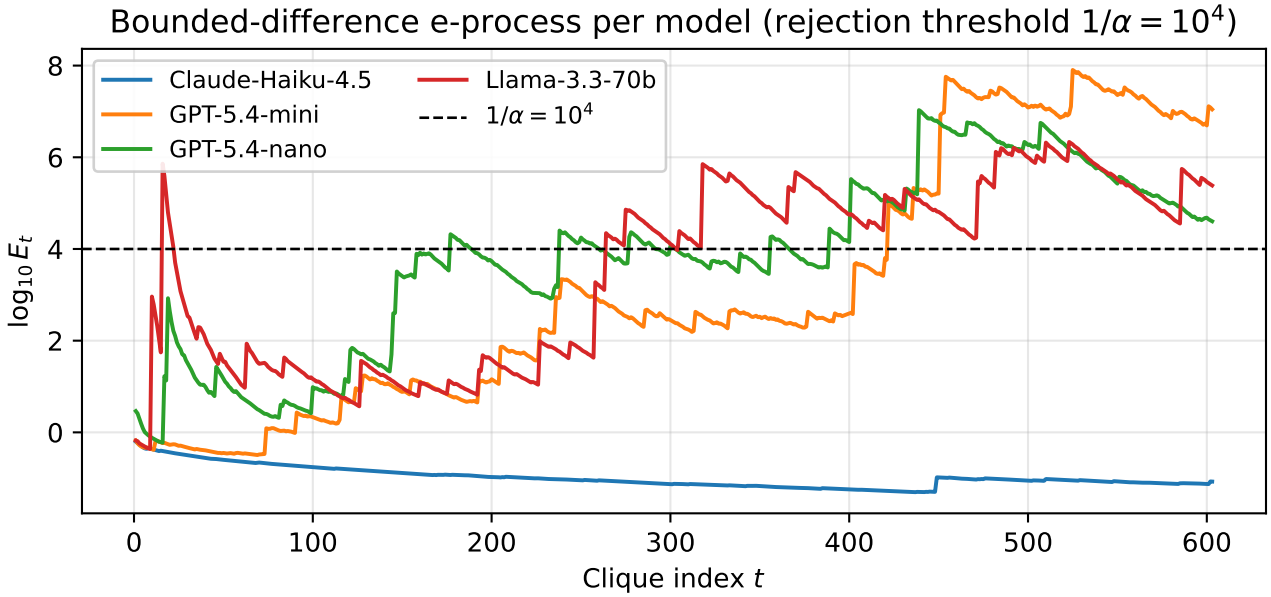

Figure 2. **Bounded-difference e-process E_t on per-component residual streams (Theorem D.2).** Crossing $1/\alpha = 10^4$ at t=16 (Llama), t=177 (GPT-nano), t=422 (GPT-5.4-mini); Claude-Haiku stays below 1.

E. Clique-mining pipeline and reproducibility

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

Polymarket. We query gamma-api at the *event* level, 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 (e.g. election winners, sports brackets); (ii) an *implication ladder* when an event has multiple threshold markets "X above k" at ascending k. Events are capped at $m \leq 8$ markets.

Compositional ensemble. For each clique with m questions and 4 LLMs, we draw 4 random seeds (independent across cliques and seeds, fixed master seed) and for each seed assign each question j to one of the 4 LLMs uniformly; the agent's quoted marginal on coordinate j is the j-th coordinate of the assigned LLM's JCD-projected forecast. Hierarchical JCD applies $\Pi^\star$ to the resulting concatenated vector (in our experiments equal to direct JCD on the joint clique, which by Theorem 3.10 is the converged Dykstra projection through the local-then-coupling cycle).

Leakage control. Every Polymarket event in the evaluation slice resolved *after* the latest model snapshot (30 April 2026). Paleka cliques inherit Paleka et al. (2025)'s own resolution-status verification.

Reproducibility manifest. Table 2 records the exact configuration used to produce every number in the paper. The released artifact contains the sample dumps, per-clique JCD residuals, ensemble-assignment seeds, and the hierarchical-projection pipeline.

Table 2. Reproducibility manifest. N values include ensemble-assignment seeds where applicable.

Item	Value
Anthropic model ID	`claude-haiku-4-5-20251001`
OpenAI model IDs	`gpt-5.4-mini`, `gpt-5.4-nano`
Meta model ID	`llama-3.3-70b-versatile` (Groq)
Snapshot date	30 April 2026
K samples / question	8
Temperature	0.7
Sampling prompts	released JSONL artifact
Paleka subset	`scraped`, 134 cliques per checker
Paleka post-filter	134×4 checkers $= 536$ cliques per model
Polymarket cliques	268 post-leakage-filter (67×4 seeds)
Per-call Platt fit / eval split	30% / 70% per (model, checker)
QP solver	OSQP default tolerance (1.4×10^{-5})
Compositional master seed	`0` (numpy default RNG)
Random ensemble seeds	4 (N=1,876 bets)
Structured ensemble seeds	1 deterministic (N=469 bets)

F. Polymarket cross-platform replication (per-call layer)

We replicate the headline per-call finding on Polymarket (N=268 held-out (model, clique) evaluations, summed across the four-model panel). Figure 3 shows the same incoherence-conditional gain shape as the Paleka result, with the $m \geq 4$ stratum (large multi-candidate partitions) contributing the largest gains. Table 3 gives the per-model breakdown.

Table 4. Real-LLM Brier (↓) on N=2412 (model, clique) evaluations across four logical-relation checkers. $\Delta_{\mathrm{B2}\to\mathrm{JCD}}$ isolates the projection's effect; $\Delta_{\mathrm{B2}\to\mathrm{B6}}$ shows the Platt+JCD combination. Negative is better.

Model	B2	B1	JCD	B5	B6	$\Delta_{\mathrm{B2}\to\mathrm{JCD}}$	$\Delta_{\mathrm{B2}\to\mathrm{B6}}$
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-versatile	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

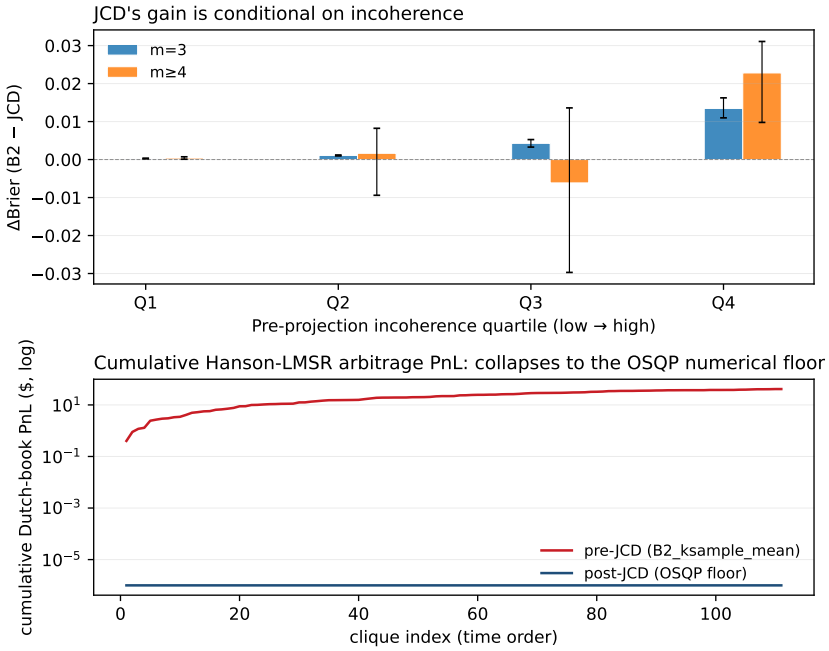

Figure 3. Polymarket cross-platform replication of the per-call result (Theorem A.1): the incoherence-conditional gain pattern of Theorem A.1 replicates on resolved Polymarket events, with the largest mean gain in the highest-ε quartile at $m\geq 4$ (note the smaller per-call absolute magnitudes; the lower panel's cumulative Hanson–LMSR exposure reduction to the QP floor matches Corollary A.2).

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. $\Delta_{\mathrm{B2}\to\mathrm{JCD}}$ isolates the projection's effect; $\Delta_{\mathrm{B2}\to\mathrm{B6}}$ shows the Platt+JCD combination. Negative is better.

Model	B2	B1	JCD	B5	B6	$\Delta_{\mathrm{B2}\to\mathrm{JCD}}$	$\Delta_{\mathrm{B2}\to\mathrm{B6}}$
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

G. Full per-cell tables

Per-model Brier under each baseline (Table 4), the six-method ablation including B3/B4 (Table 5), the significance tests (Table 6), and the Murphy reliability/resolution decomposition (Table 7).

Table 5. 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 6. Significance tests for H_0: mean Brier(B2) = mean Brier(JCD).

Model	DM stat	DM p	paired-boot stat	paired-boot p
anthropic_claude-haiku-4-5-20251001	5.422	8.52e-08	0.0066	0
openai_gpt-5.4-mini	10.682	1.657e-24	0.0150	0
openai_gpt-5.4-nano	12.281	4.277e-31	0.0148	0
meta_llama-3.3-70b-versatile	7.980	7.437e-15	0.0079	0

H. Proof of Cor. 3.9 (Rayleigh-quotient magnitude bound)

We give the proof under uniform i.i.d. owner-selection $\sigma : [m^\star] \to [k]$, with $\bar{\Pi} := k^{-1} \sum_a \Pi_a$ and $D := \mathrm{diag}(\Sigma_\Pi)$. Define $u := x - \bar{\Pi}$ where $x_j := \Pi_{\sigma(j),j}$. Note $\bar{\Pi} \in \mathcal{M}^\star$ by convexity of $\mathcal{M}^\star$, and pointwise $\mathbb{E}_\sigma[u_j] = 0$, $\mathbb{E}_\sigma[u_j^2] = D_{jj}$, and cross-coordinate covariances vanish: $\mathbb{E}_\sigma[u_j u_{j'}] = 0$ for $j \neq j'$ by independence of $\sigma(j), \sigma(j')$.

Generic upper bound. By Prop. 3.8 with reference $r = \bar{\Pi}$, $\varepsilon^\star \leq \|u\|_2$, so $\mathbb{E}_\sigma[(\varepsilon^\star)^2] \leq \mathbb{E}_\sigma[\|u\|_2^2] = \sum_j D_{jj} = \mathrm{tr}(D)$.

Equality case ($\kappa_R = 1$). If $\mathcal{M}^\star$ contains a single equality $a_R^\top r = b_R$ binding at the projection (and no other constraint is active locally), the projection has the closed form $\Pi^\star(x) = x - \frac{a_R^\top x - b_R}{\|a_R\|^2} a_R$, giving $\varepsilon^\star = |a_R^\top x - b_R| / \|a_R\|$. Since $\bar{\Pi} \in \mathcal{M}^\star$ satisfies $a_R^\top \bar{\Pi} = b_R$, we have $a_R^\top x - b_R = a_R^\top u$, so $\varepsilon^{\star 2} = (a_R^\top u)^2 / \|a_R\|^2$. Taking expectation, $\mathbb{E}_\sigma[(a_R^\top u)^2] = a_R^\top D a_R$ (cross-terms vanish), so $\mathbb{E}_\sigma[(\varepsilon^\star)^2] = a_R^\top D a_R / \|a_R\|^2$.

Inequality case with boundary $\bar{\Pi}$ ($\kappa_R = \frac{1}{2}$). If the binding constraint at the projection is an inequality $a_R^\top r \leq b_R$ with $a_R^\top \bar{\Pi} = b_R$ (boundary), the single-halfspace projection gives $\varepsilon^\star = (a_R^\top u)_+ / \|a_R\|$. The variable $a_R^\top u$ is zero-mean and symmetrically distributed about 0 when the panel $\{\Pi_a\}$ is symmetric in $a_R^\top \Pi_a$ around its mean b_R (which holds asymptotically and approximately holds for our $k = 4$ panel). Then $\mathbb{E}[(\cdot)_+^2] = \frac{1}{2}\mathbb{E}[(\cdot)^2]$, giving $\mathbb{E}_\sigma[(\varepsilon^\star)^2] = \frac{1}{2} \cdot a_R^\top D a_R / \|a_R\|^2$.

Inequality case with interior $\bar{\Pi}$. When $\bar{\Pi}$ is strictly interior to constraint i (slack $s_i := b_i - a_i^\top \bar{\Pi} > 0$), the violation $v_i = (a_i^\top u - s_i)_+$ has reduced second moment under the truncation, so $\mathbb{E}_\sigma[(\varepsilon^\star)^2] < \frac{1}{2} a_R^\top D a_R / \|a_R\|^2$ in this regime; the bound becomes a strict upper bound. This matches the empirical conjunction under-shoot (0.83× predicted): $\bar{\Pi}_3 < \min(\bar{\Pi}_1, \bar{\Pi}_2)$ by Jensen, leaving the upper Fréchet constraint slack at $\bar{\Pi}$.

Closed-form residuals. The negation case (equality, $a = (1, 1)$, $b = 1$, $\|a\|^2 = 2$) recovers $\mathbb{E}[(\varepsilon^\star)^2] = \mathrm{tr}(D)/2$. The partition case (equality, $a = \mathbf{1}_{m^\star}$, $b = 1$, $\|a\|^2 = m^\star$) recovers $\mathrm{tr}(D)/m^\star$ in the no-clipping regime; with clipping, the actual residual is strictly larger, leaving this as a lower bound. The conjunction/disjunction cases use Fréchet halfspace normals as a_R (with $\|a_R\|^2 \in \{2, 3\}$ depending on the binding constraint per clique), with $\kappa_R = \frac{1}{2}$. □

Empirical validation. Table 8 reports observed $\langle(\varepsilon^\star)^2\rangle$ versus the Rayleigh-quotient prediction $\kappa_R \, a_R^\top D a_R / \|a_R\|^2$ across 1,876 ensemble cliques. The dominant active constraint a_R per clique is the defining halfspace with smallest slack at $\bar{\Pi}$ (typically the upper Fréchet for and, the upper Fréchet $r_3 \leq r_1 + r_2$ for or).

Table 7. 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 8. Predicted vs. observed $\langle(\varepsilon^\star)^2\rangle$ under uniform random owner-selection. "Generic" is $\mathrm{tr}(D)$ (the upper bound without single-active-constraint geometry); "Rayleigh" is $\kappa_R\, a_R^\top D a_R/\|a_R\|^2$.

Relation	κ_R	$\|a_R\|^2$	$\langle$obs$\rangle$	$\langle$Rayleigh$\rangle$	ratio	per-clique r
NEG	1	2	0.0286	0.0271	1.054	0.890
PARTITION	1	$m^\star$	0.0146	0.0137	1.069	0.850
AND	$\frac{1}{2}$	$\{2,3\}$	0.0153	0.0184	0.830	0.710
OR	$\frac{1}{2}$	$\{2,3\}$	0.0170	0.0166	1.026	0.690

The Rayleigh form is uniformly tight: ratios are 0.83–1.07 across all four relations. The generic $\mathrm{tr}(D)$ bound is loose by $1.9\times$ on NEG, $4.0\times$ on PARTITION (at the average $m^\star = 4$), and 5.3–$6.1\times$ on conjunction/disjunction; the Rayleigh form recovers the typical magnitude from the specialist panel covariance alone.

I. Compositional ablation: projection-step decomposition

We decompose hierarchical JCD into two layers (per-component JCD then joint projection) and four operators evaluated against resolved labels: (A) raw composed (no per-component JCD, no joint projection), (B) JCD composed (per-component JCD only), (C) raw + joint projection, (D) JCD + joint projection (hierarchical). Table 9 reports mean $\varepsilon^\star$ before projection (A, B) and mean compositional Brier under each operator.

Table 9. Compositional projection-step ablation. $\varepsilon^\star$ is post-aggregation; Brier is against resolved labels. Hier $\equiv$ D.

Relation	Op	$\langle\varepsilon^\star\rangle$	exposure	Brier	ΔBrier vs. A
NEG	A: raw	0.144	0.203	0.2367	
	B: JCD	0.114	0.161	0.2301	−0.0066
	C: raw $+\Pi^\star$	0	0	0.2174	−0.0193
	D: hier	0	0	0.2165	−0.0202
AND	A	0.076	0.110	0.2270	
	B	0.054	0.079	0.2186	−0.0084
	C	0	0	0.2144	−0.0126
	D	0	0	0.2110	−0.0160
OR	A	0.110	0.153	0.2538	
	B	0.072	0.102	0.2557	+0.0019
	C	0	0	0.2571	+0.0033
	D	0	0	0.2584	+0.0046
PARTITION	A	0.312	0.620	0.2165	
	B	0.098	0.193	0.1864	−0.0301
	C	0	0	0.1832	−0.0333
	D	0	0	0.1817	−0.0348

Per-component JCD dominates the partition gain (the local repair already brings the assembled mass close to 1). The joint projection dominates negation, and, and disjunction, where local repair leaves cross-coordinate violations untouched. The disjunction inversion under (B)–(D) is the artefact discussed in §4.

J. Greedy-decoding ($T=0$) same-model control

To isolate within-model decoding stochasticity from coordinate-isolation disagreement, we re-ran the 4-fresh-seed Claude-Haiku-4.5 same-model protocol of §4 at temperature $T=0$ on 60 cliques per Paleka relation and 30 Polymarket partitions. Numbers in Table 10 are mean $\varepsilon^\star$ under 16 uniform random owner-selection draws per clique; the $T=0.7$ row is the sampling-temperature baseline of §4.

Table 10. Same-model $\varepsilon^\star$ at $T=0.7$ vs. $T=0$ (greedy decoding). Cross-model column reproduced from §4 for reference.

Relation	cross-model	same-model $T=0.7$	same-model $T=0$	$T_0/T_{0.7}$	T_0/cross
NEG	0.114	0.025	0.096	3.83	0.84
AND	0.054	0.019	0.023	1.20	0.42
OR	0.072	0.040	0.074	1.87	1.04
PARTITION	0.098	0.058	0.025	0.44	0.26

The greedy-decoding control rules out the alternative "$\varepsilon^\star$ is just sampling noise" hypothesis. On negation and disjunction, $T=0$ *increases* the residual ($T_0/T_{0.7}>1$); on partition it decreases it by 56%, indicating that some partition residual is decoding-stochastic. On all four relations the residual remains positive at $T=0$ and is comparable in magnitude to (or, for negation/disjunction, larger than) the $T=0.7$ same-model figure. The result implicates genuine coordinate-isolation disagreement: the same model, prompted with isolated single-question contexts at greedy decoding, produces different "deterministic" outputs across coordinates that fail to satisfy the cross-component coupling. (Anthropic does not accept a `seed` parameter, so 4 "specialists" at $T=0$ vary only through residual non-determinism in batch processing; $\varepsilon^\star$ remains detectable nonetheless.)

K. K-sweep on real LLM forecasts

Per-component layer. Figure 4 sweeps $K \in \{4, 8, 16, 24\}$ on Anthropic Claude-Haiku-4.5 NegChecker forecasts. The non-zero ΔBrier floor (0.008–0.010) is the empirical signature of $p_F \notin \mathcal{M}_C$ rather than finite-sample fluctuation.

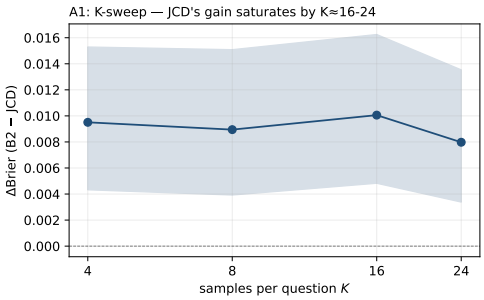

Figure 4. K-sweep on Anthropic Claude-Haiku-4.5 NegChecker forecasts. Non-vanishing ΔBrier asymptote: the empirical signature of structural incoherence in p_F.

Compositional layer. We also sweep $K \in \{8, 16, 32\}$ at the compositional layer (the residual $\varepsilon^\star$ in §4). $K=8$ uses the existing 4-LLM panel and $K=16, 32$ extends Anthropic Claude-Haiku-4.5, GPT-5.4-mini, and GPT-5.4-nano to additional samples on a 40-clique-per-relation, 20-partition subset (Groq Llama-3.3-70b excluded after exhausting its tokens-per-day quota mid-sweep). The 3-model panel ensemble is computed identically to the 4-LLM panel under uniform random owner-selection, 16 ensemble draws per clique.

Table 11. Compositional $\varepsilon^\star$ vs. K, 3-model panel.

Relation	$K=8$	$K=16$	$K=32$	N
NEG	0.143	0.145	0.143	432
AND	0.067	0.067	0.067	432
OR	0.078	0.080	0.079	432
PARTITION	0.076	0.082	0.082	224

The compositional residual is flat across $K \in \{8, 16, 32\}$ (within ± 0.003 per relation), confirming that $\varepsilon^\star$ is a *structural* property of cross-component coupling under specialist routing rather than finite-sample fluctuation. Combined with the per-component flatness above, the K-sweep robustness covers both layers.

L. Structured-routing replication

Figure 5 repeats the compositional experiment under the structured-routing topology of Section 4.

The residual is positive on 98.5% of negations (vs. 66% under random), 36% of partitions (vs. 94%), 12% of conjunctions (vs. 33%), and 16% of disjunctions (vs. 43%). Per-bet exposure is \$0.11/bet (\$51.88 over 469 bets), comparable to the random regime's \$0.13/bet. Negation has a higher residual rate because two specialists disagree more strongly than four random ones; conjunction/disjunction improve because structured routing reduces cross-model mixing, while partitions improve because one specialist owns all candidates.

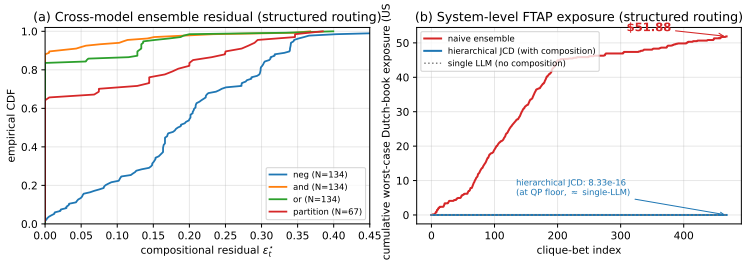

Figure 5. **Structured-routing replication of Figure 1.** The same certificate is positive with a different relation breakdown but comparable per-bet exposure (\$0.11/bet vs. \$0.13/bet under random assignment). The lower cumulative endpoint reflects fewer bets under structured routing, N=469 vs. N=1,876, not lower per-bet exposure.

M. Worst-case failure modes

Table 12 lists the top-5 ensemble cliques by compositional residual $\varepsilon^\star$ under random assignment. The largest cases are realized resolved-Paleka quotes where specialists make opposite sides of a negation jointly too likely, or make a disjunction inconsistent with its antecedents.

Table 12. Top-5 ensemble cliques by compositional residual $\varepsilon^\star$ under random assignment (N=1,876).

#	relation	m	component quotes (assigned LLM)	violated constraint	$\varepsilon^\star$
1	neg	2	0.84 (Claude), 0.89 (Llama)	$p_1 + p_2 = 1.731 \neq 1$	0.517
2	or	3	0.00 (Claude), 0.02 (GPT), 0.92 (Claude)	$p_3 = 0.916 < \max(p_1, p_2) = 0.020$	0.517
3	or	3	0.02 (Claude), 0.03 (GPT), 0.92 (Claude)	$p_3 = 0.924 < \max(p_1, p_2) = 0.031$	0.502
4	or	3	0.02 (Claude), 0.74 (GPT), 0.03 (Claude)	$p_3 = 0.033 < \max(p_1, p_2) = 0.740$	0.500
5	neg	2	0.71 (GPT), 0.99 (Claude)	$p_1 + p_2 = 1.696 \neq 1$	0.492

N. Planner-style routing simulation: per-question detail

We report all 100 live-style partition forecasting questions used in the planner-style routing simulation (Section 4). Each row gives the partition label, the number of outcomes m, the per-outcome marginal returned by the randomly-assigned specialist LLM (LLM key: `C`=Claude-Haiku-4.5, `M`=GPT-5.4-mini, `N`=GPT-5.4-nano, `L`=Llama-3.3-70b), the resulting partition mass $\sum_i p_i$, and the compositional residual $\varepsilon^\star$. All four specialist clients ran with K=8 verbalized-probability samples at temperature 0.7; each received its assigned outcome as an isolated single-question prompt with no context about the partition or the other outcomes. Master seed 0, sampling without replacement when $m \leq 4$. Total cost: $\approx$2,400 LLM calls, $\approx$\$5, $\approx$35 minutes wall-clock. Rows are sorted by descending $\varepsilon^\star$ so the worst cases appear first.

Table 13. Planner-style routing simulation, all $N{=}100$ partitions, sorted by $\varepsilon^\star$. Each specialist sees a single Bernoulli question; the agent assembles into a partition quote. LLM key: C=Claude-Haiku-4.5, M=GPT-5.4-mini, N=GPT-5.4-nano, L=Llama-3.3-70b. All 100 partitions exhibit positive compositional residual.

Partition	m	per-outcome quote (LLM)	$\sum p_i$	$\varepsilon^\star$
Largest US AI startup IPO 2026	4	0.39(C), 0.73(N), 0.67(L), 0.71(M)	2.497	0.749
Top MMLU score, year-end 2026	4	0.25(C), 0.62(M), 0.57(N), 0.85(L)	2.290	0.650
2026 Nobel Chemistry laureate field	4	0.59(C), 0.38(N), 0.57(L), 0.61(M)	2.164	0.582
Top SWE-Bench Verified score, year-end 2026	4	0.35(N), 0.61(L), 0.25(C), 0.86(M)	2.079	0.540
Next UK general election year	4	0.32(N), 0.23(L), 0.72(C), 0.74(M)	2.008	0.504
2026 ACM Turing Award area	3	0.58(M), 0.54(C), 0.74(L)	1.870	0.502
Top HumanEval+ score, year-end 2026	4	0.25(C), 0.48(N), 0.43(L), 0.78(M)	1.936	0.468
Top GPQA score, year-end 2026	4	0.34(N), 0.36(C), 0.42(L), 0.80(M)	1.925	0.462
2026 EU AI Act state of enforcement	3	0.30(N), 0.74(M), 0.73(L)	1.775	0.447
2026 German Bundestag composition (post-election)	3	0.72(C), 0.42(L), 0.59(N)	1.735	0.424
2026 Nobel Economics laureate field	4	0.50(L), 0.44(N), 0.35(M), 0.53(C)	1.813	0.406
2026 Emmy Outstanding Drama Series platform	3	0.74(L), 0.67(M), 0.29(N)	1.701	0.405
Top global cloud provider revenue share 2026	4	0.72(N), 0.39(L), 0.08(C), 0.58(M)	1.768	0.403
Next iPhone announcement quarter	3	0.70(M), 0.80(L), 0.17(C)	1.676	0.395
2026 Ohio gubernatorial party	3	0.71(C), 0.40(N), 0.55(L)	1.660	0.381
2026 US House majority	3	0.44(N), 0.58(L), 0.64(M)	1.660	0.381
Top open-source LLM family year-end 2026	4	0.71(C), 0.23(M), 0.39(N), 0.43(L)	1.759	0.379
2026 International Booker Prize winner language	3	0.68(C), 0.43(M), 0.54(N)	1.651	0.376
2027 UK Prime Minister tenure	3	0.52(C), 0.41(M), 0.71(L)	1.649	0.375
2026 Wolf Prize in Mathematics	3	0.77(M), 0.35(C), 0.50(L)	1.618	0.357
UN climate ambition state, COP31 (2026)	3	0.59(N), 0.71(C), 0.29(M)	1.591	0.341
Apple WWDC 2026 keynote	3	0.54(N), 0.79(C), 0.25(M)	1.579	0.334
Largest US tech IPO of 2026 by company type	4	0.43(L), 0.36(C), 0.44(M), 0.43(N)	1.665	0.333
ITER first plasma timing	3	0.34(N), 0.50(L), 0.73(C)	1.573	0.331
2026 Nobel Physics laureate field	4	0.43(C), 0.44(L), 0.30(M), 0.47(N)	1.636	0.318
2027 Hugo Award Best Novel author origin	3	0.70(L), 0.17(N), 0.65(M)	1.529	0.305
2027 ICC T20 Cricket World Cup winner region	3	0.55(C), 0.42(N), 0.55(L)	1.526	0.304
2026 Cannes Palme d'Or winner nationality	3	0.57(C), 0.27(M), 0.68(L)	1.518	0.299
2026 H1 BTC return bucket	3	0.42(L), 0.38(N), 0.71(C)	1.515	0.297
2026 Nobel Medicine laureate field	4	0.20(N), 0.36(M), 0.28(C), 0.75(L)	1.589	0.294
2026 Grammy Album of the Year genre	4	0.52(M), 0.28(C), 0.36(L), 0.42(N)	1.577	0.289
2026 FIFA World Cup winner confederation	3	0.56(C), 0.42(N), 0.52(L)	1.498	0.287
Top semiconductor maker by revenue, 2026	3	0.57(M), 0.57(C), 0.35(N)	1.494	0.285
US unemployment rate, June 2026	4	0.71(C), 0.17(N), 0.28(L), 0.38(M)	1.540	0.270
ECB main refinancing rate, December 2026	3	0.49(N), 0.55(M), 0.42(L)	1.460	0.266
2027 Six Nations rugby winner	3	0.39(M), 0.64(C), 0.42(L)	1.450	0.260
2027 Mexico midterm legislative balance	3	0.72(C), 0.28(N), 0.44(L)	1.442	0.255
2026 US CPI year-over-year, December print	4	0.30(N), 0.46(M), 0.42(L), 0.33(C)	1.507	0.254
2026 US real GDP year-over-year growth	4	0.34(N), 0.56(L), 0.26(M), 0.34(C)	1.499	0.249
2027 French Presidential winner ideology	3	0.33(C), 0.52(M), 0.58(L)	1.430	0.248
Next FOMC decision (June 2026 meeting)	3	0.33(N), 0.68(C), 0.42(L)	1.424	0.245

(continued on next page)

(continued from previous page)

Partition	m	per-outcome quote (LLM)	$\sum p_i$	$\varepsilon^\star$
2026 Nobel Literature laureate region	3	0.62(C), 0.35(N), 0.43(L)	1.399	0.230
2026 IPhO top country	3	0.24(N), 0.42(L), 0.73(C)	1.394	0.227
Top semiconductor revenue 2026: foundry vs IDM	3	0.25(C), 0.50(L), 0.63(N)	1.380	0.219
2026 Tour de France winner nationality	3	0.26(L), 0.40(N), 0.72(C)	1.377	0.218
December 2026 FOMC decision	3	0.49(M), 0.46(C), 0.42(L)	1.371	0.214
2027 Australian Open men's singles winner	4	0.10(C), 0.25(M), 0.35(L), 0.72(N)	1.419	0.209
2027 Breakthrough Prize Life Sciences focus	4	0.27(N), 0.35(C), 0.50(L), 0.30(M)	1.417	0.209
2026 Academy Award Best Picture winner type	3	0.72(C), 0.43(L), 0.20(N)	1.353	0.204
2027 BAFTA Best Film origin	3	0.36(C), 0.71(L), 0.29(M)	1.351	0.203
2026 New York gubernatorial party	3	0.83(M), 0.26(N), 0.24(C)	1.336	0.194
2026 NCAA Division I football national champion conference	3	0.44(N), 0.28(C), 0.60(L)	1.321	0.185
2026 Brazilian general election composition	3	0.57(N), 0.32(C), 0.42(L)	1.309	0.178
June 2027 FOMC decision	3	0.42(L), 0.66(C), 0.21(M)	1.295	0.170
WHO PHEIC declarations 2026	3	0.16(N), 0.44(C), 0.11(M)	0.706	0.170
2026 F1 Constructors' Championship winner	4	0.27(N), 0.38(L), 0.18(M), 0.50(C)	1.336	0.168
UK CPI YoY November 2026	3	0.54(L), 0.31(N), 0.43(M)	1.276	0.159
2026 Mercury Music Prize winner gender	3	0.47(C), 0.40(L), 0.40(M)	1.268	0.154
2026 H1 SP500 return bucket	4	0.35(N), 0.50(L), 0.21(C), 0.25(M)	1.302	0.151
2026 Nobel Peace Prize laureate type	3	0.75(N), 0.44(M), 0.06(C)	1.252	0.148
2026 Tony Best Musical genre	3	0.35(C), 0.27(M), 0.63(N)	1.254	0.147
World population, mid-2027	3	0.22(M), 0.42(C), 0.60(N)	1.245	0.141
2026 F1 Drivers' Championship winner	4	0.29(M), 0.26(N), 0.35(L), 0.36(C)	1.259	0.129
EUR/USD year-end 2026	3	0.21(M), 0.69(C), 0.29(N)	1.204	0.118
Gold price year-end 2026	3	0.35(C), 0.12(N), 0.74(M)	1.204	0.118
WTI crude oil year-end 2026 price bucket	3	0.45(M), 0.43(L), 0.31(N)	1.199	0.115
2026 Pennsylvania US Senate election	3	0.45(C), 0.48(L), 0.24(N)	1.175	0.101
Tesla 2026 Q4 deliveries bucket	3	0.35(C), 0.50(L), 0.32(M)	1.170	0.098
2026 Atlantic hurricane season activity	3	0.39(N), 0.47(M), 0.30(C)	1.165	0.095
2027 Berlin Golden Bear winner type	3	0.69(N), 0.09(M), 0.07(L)	0.847	0.088
2027 H1 SP500 return bucket	3	0.09(N), 0.34(M), 0.43(L)	0.848	0.088
2026 Wimbledon men's singles winner	4	0.12(C), 0.08(N), 0.12(M), 0.85(L)	1.172	0.086
2026 Florida gubernatorial party	3	0.58(L), 0.25(C), 0.31(N)	1.140	0.081
US 10-year Treasury yield year-end 2026	3	0.30(M), 0.42(C), 0.42(L)	1.139	0.080
2026 Olivier Best New Play origin	3	0.70(C), 0.38(M), 0.05(L)	1.136	0.079
2026 IMO top country	3	0.36(N), 0.27(M), 0.50(L)	1.134	0.077
2026 mpox global status	3	0.49(M), 0.13(N), 0.25(C)	0.867	0.076
2026 H2 SP500 return bucket	4	0.12(N), 0.29(C), 0.28(L), 0.17(M)	0.859	0.071
2026 global temperature ranking	3	0.27(N), 0.11(M), 0.50(L)	0.883	0.068
2026 PGA Tour FedEx Cup winner nationality	2	0.68(M), 0.23(N)	0.905	0.067
Cancer screening trial readout 2026	3	0.23(M), 0.66(C), 0.23(L)	1.114	0.066
2026 Texas gubernatorial party	3	0.55(N), 0.16(C), 0.18(M)	0.895	0.061
2026 US Senate majority	3	0.58(L), 0.27(M), 0.26(N)	1.104	0.060
Largest 2026 US data-center investment site	3	0.37(M), 0.23(N), 0.50(L)	1.098	0.056
2026 Arctic sea ice minimum extent	3	0.62(C), 0.13(N), 0.34(M)	1.095	0.055
2026 Japan LDP leadership	3	0.67(L), 0.33(N), 0.09(C)	1.090	0.052
2026 ACM Gödel Prize area	3	0.43(C), 0.15(M), 0.50(L)	1.085	0.049

(continued on next page)

(continued from previous page)

Partition	m	per-outcome quote (LLM)	$\sum p_i$	$\varepsilon^\star$
2026 NBA Finals winner conference	2	0.53(L), 0.53(M)	1.062	0.044
March 2027 FOMC decision	3	0.35(C), 0.29(N), 0.42(L)	1.059	0.034
2026 NHL Stanley Cup winner conference	2	0.50(C), 0.54(M)	1.035	0.025
2026 World Series winner league	2	0.53(M), 0.51(N)	1.034	0.024
2026 California gubernatorial party	3	0.50(N), 0.16(C), 0.38(L)	1.038	0.022
2026 net global solar capacity addition	3	0.20(C), 0.44(M), 0.33(N)	0.965	0.020
Top gold-medal nation, 2028 Summer Olympics	3	0.58(L), 0.19(M), 0.26(N)	1.031	0.018
2028 US Presidential winner	3	0.46(L), 0.53(M), 0.02(C)	1.016	0.009
Year-end 2026 leading mobile OS	3	0.85(L), 0.15(M), 0.02(C)	1.016	0.009
GLP-1 agonist label expansion 2026	3	0.30(N), 0.57(M), 0.11(C)	0.984	0.009
FDA Alzheimer's drug action 2026	3	0.33(M), 0.18(N), 0.50(L)	1.012	0.007
2026 Booker Prize winner nationality	3	0.37(M), 0.18(N), 0.44(L)	0.996	0.002
2026 Pulitzer Prize Fiction winner setting	3	0.72(C), 0.25(M), 0.03(N)	0.999	0.001

Worst-case anatomy. The largest residual ($\varepsilon^\star = 0.749$) is the "Largest US AI startup IPO 2026" partition (four sectoral outcomes: infrastructure, model, applications, other). Three of the four assigned specialists privately quote $P(\text{their sector}) > 0.5$ (the fourth is at 0.39); collectively the specialists over-allocate, with assigned masses summing to 2.50. Each marginal is locally well-calibrated for the single Bernoulli the specialist was asked, but no specialist saw that the four sectors tile the field; as a result, the assembled system quotes 250% probability mass. Simplex projection reduces the quote to a proper distribution in closed form, eliminating the exposure to QP-tolerance. ML-benchmark partitions (Top MMLU, $\varepsilon^\star$=0.65; SWE-Bench Verified, 0.54; HumanEval+, 0.47; GPQA, 0.46) form a coherent high-residual cluster: in each, four specialists are assigned to one candidate model organization, and the specialists collectively over-allocate; masses sum to ~ 2.0 even when only one or two specialists are above 0.5 individually.

Why some partitions yield near-zero residual. The five smallest residuals (Pulitzer Prize Fiction setting 0.001, Booker Prize nationality 0.002, FDA Alzheimer's drug action 0.007, GLP-1 label expansion 0.009, leading mobile OS 0.009) share a common structure: the partition either has only two real candidate outcomes, or has a strongly bimodal training-set prior. The single-question marginals then approximately satisfy the partition constraint without any explicit cross-component coordination. The certificate is therefore discriminating rather than uniformly positive: it is positive on novel multi-way partitions and near zero on well-rehearsed ones.

O. Routing-protocol ladder: per-partition detail

We compare seven routing protocols on the same 100 live-style multi-candidate partition benchmark (App. N). Three protocols use fixed owner-selected routing under different prompt templates (ISOLATED, LISTED, FULL); four use a tool-using planner (Claude-Haiku-4.5 or GPT-5.5 as planner, each $\times$ unguided/coherence-guided framings). Specialist panel across all protocols: Claude-Haiku-4.5, GPT-5.4-mini, GPT-5.4-nano, DeepSeek-V3.2, Llama-4-Maverick-17B-128E (5 specialists; Llama-3.3-70b dropped after TPD-quota issues). $K = 8$ verbalized samples per specialist at temperature 0.7. Total cost: $\approx$\$80 across all protocols.

Prompt templates. ISOLATED reuses the per-question `DEFAULT_PROMPT` of Paleka et al. (2025)-style routing. LISTED prepends a numbered list of all m partition outcomes followed by "Specific outcome to estimate: {*outcome*}"; no normalization instruction. FULL adds the explicit clause: "these outcomes are mutually exclusive and collectively exhaustive, so the probabilities you would assign to all m outcomes must sum to 1. Bear this in mind."

Table 14. Routing-protocol ladder ($N = 100$ partitions). Cost is mean delegate calls per partition divided by mean partition size $\bar{m} \approx 3.3$. CI bracketed.

protocol	$\langle\varepsilon^\star\rangle$	median	frac. $\varepsilon^\star > 10^{-3}$	cost
Owner-selected routing, prompt-engineering only:				
ISOLATED	$0.235\ [0.20, 0.27]$	0.186	1.00	$1\times m$
LISTED	$0.090\ [0.08, 0.11]$	0.077	0.98	$1\times m$
FULL+SUM-TO-1	$0.081\ [0.07, 0.09]$	0.070	0.99	$1\times m$
Tool-using planner with delegate + submit-final tools:				
Claude-Haiku, unguided	$0.0090\ [0.001, 0.024]$	0.000	0.05	$4.69\times m$
Claude-Haiku, guided	$0.0044\ [0.000, 0.011]$	0.000	0.03	$4.87\times m$
GPT-5.5, unguided	$0.0000\ [0,\ 0]$	0.000	0.00	$6.98\times m$
GPT-5.5, guided	$0.0000\ [0,\ 0]$	0.000	0.00	$6.51\times m$
Geometric repair (any prompt, owner-selected):				
hierarchical JCD	$\leq 10^{-7}$ (QP floor)	0	0	$1\times m + \Pi^\star$

Pairwise Wilcoxon: prompt ISO $\rightarrow$ LISTED $p = 2.2\times 10^{-12}$, but LISTED $\rightarrow$ FULL $p = 0.18$ (showing alternatives matters; sum-to-1 instruction does not). Planner Claude-Haiku $>$ GPT-5.5: $p < 10^{-3}$ (unguided), $p < 10^{-4}$ (guided): reasoning depth strictly closes the residual that tool-use alone leaves.

Where context helps most. The largest drops (ISO $\rightarrow$ FULL) are on partitions where isolated specialists strongly anchor to their assigned outcome with no cross-context knowledge of alternatives: "Largest US AI startup IPO 2026" ($0.74 \rightarrow 0.12$), "Top MMLU score, year-end 2026" ($0.60 \rightarrow 0.05$), "2026 ACM Turing Award area" ($0.63 \rightarrow 0.13$). These are partitions over emerging or specialist-knowledge candidates where the specialist needs to know the alternatives to allocate probability mass appropriately.

Where context hurts. Eighteen partitions show $\varepsilon^\star_{\text{full}} > \varepsilon^\star_{\text{iso}}$, predominantly political/economic ones where isolated specialists were already roughly normalised (e.g., 2028 US Presidential winner $0.003 \rightarrow 0.022$, 2026 California gubernatorial party $0.10 \rightarrow 0.26$). When isolated estimates already average to ~ 1, forcing specialists to reason about all alternatives can introduce mass that wasn't there.

Implication for deployment. Three regimes emerge: (i) *prompt-engineering at fixed owner-selection* attenuates the failure mode but bottoms out at $\varepsilon^\star \approx 0.08$ (essentially the ISOLATED $\rightarrow$ LISTED step); (ii) *tool-using planners* reach near-zero residual but pay 5–7$\times$ the naive elicitation budget, with frontier reasoning depth required to close the last ~ 0.005 that tool-use alone leaves; (iii) *hierarchical projection* reaches the same near-zero residual at the naive $1\times m$ elicitation cost, modulo a small QP solve. The geometric repair is therefore the cost-efficient route to coherence when the coupling set $\mathcal{C}$ is specifiable; smart-planner integration is the alternative when $\mathcal{C}$ is implicit and the planner can infer it through extra delegate calls.

P. Tool-augmented specialist A/B: per-partition detail

We re-run the first 30 partitions of the case study under a tool-augmented protocol (§4). Each specialist issues exactly one DuckDuckGo `web_search` call against its assigned outcome text, and the top-five title+snippet results are inserted as context before the specialist emits its $K{=}8$ verbalized-probability samples. Master seed and assignment are identical to the no-tool first-30 baseline, so the per-partition $\Delta\varepsilon^\star = \varepsilon^\star_t - \varepsilon^\star_b$ is a direct A/B. Total cost: $\approx$720 LLM calls + 90 search calls, $\approx$\$2, $\approx$25 minutes wall-clock.

Aggregate: mean $\varepsilon^\star_b = 0.260 \rightarrow \varepsilon^\star_t = 0.283$; $20/30$ partitions are equal or worse with retrieval, $10/30$ are better; mean $\Delta\varepsilon^\star = +0.023$. The single largest *worsening* is "World population, mid-2027" ($+0.563$): every specialist's search returns recent UN bracket estimates around 8.1–8.2 bn and each anchors near $P(\text{their bucket}) \approx 0.7$, summing to 2.22. The single largest *improvement* is "2026 Grammy Album of the Year genre" (-0.267): retrieval surfaces 2026 Grammy nominee distributions that pull every specialist toward a similar partition shape.

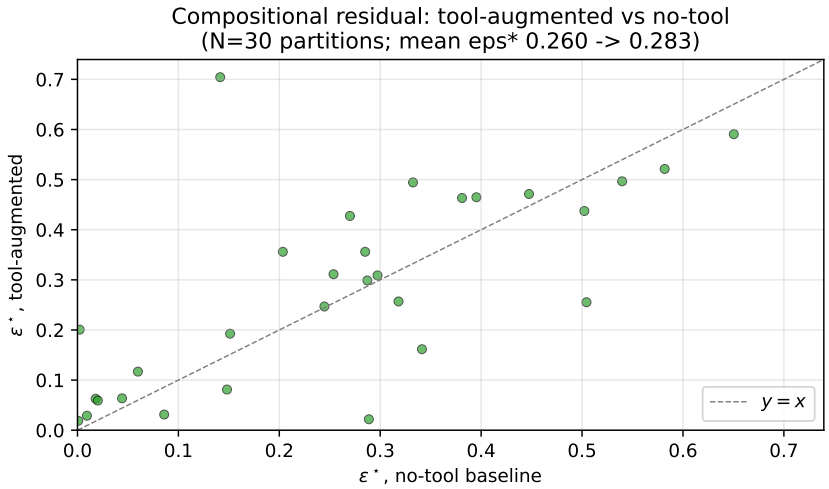

Figure 6. **Tool augmentation does not eliminate the compositional residual.** Each point is one of the 30 matched partitions; x-axis is $\varepsilon^\star$ in the no-tool baseline, y-axis is $\varepsilon^\star$ when each specialist receives top-five DuckDuckGo snippets for its assigned outcome before quoting. Most points lie on or above $y{=}x$: tool augmentation tightens within-specialist variance but cross-specialist disagreement persists, so the assembled partition quote is no more coherent with retrieval than without.

Table 15. Tool-augmented specialist A/B, all 30 matched partitions, sorted by $|\Delta\varepsilon^\star|$. $\sum_t$ is the partition mass under tool-augmented retrieval; $\varepsilon_b^\star$ and $\varepsilon_t^\star$ are the no-tool baseline and tool-augmented compositional residuals.

Partition	$\sum_t$	$\varepsilon_b^\star$	$\varepsilon_t^\star$	$\Delta\varepsilon^\star$
World population, mid-2027	2.220	0.141	0.704	+0.563 ↑
2026 Grammy Album of the Year genre	1.044	0.289	0.022	-0.267 ↓
Next UK general election year	0.489	0.504	0.255	-0.249 ↓
2026 Booker Prize winner nationality	1.348	0.002	0.201	+0.198 ↑
UN climate ambition state, COP31 (2026)	1.280	0.341	0.162	-0.180 ↓
Largest US tech IPO of 2026 by company type	1.989	0.333	0.494	+0.162 ↑
US unemployment rate, June 2026	1.855	0.270	0.427	+0.158 ↑
2026 Academy Award Best Picture winner type	1.616	0.204	0.356	+0.152 ↑
2026 US House majority	1.802	0.381	0.463	+0.082 ↑
Top semiconductor maker by revenue, 2026	1.566	0.285	0.356	+0.071 ↑
Next iPhone announcement quarter	1.805	0.395	0.465	+0.069 ↑
2026 Nobel Peace Prize laureate type	1.131	0.148	0.081	-0.067 ↓
2026 ACM Turing Award area	1.757	0.502	0.437	-0.065 ↓
2026 Nobel Physics laureate field	1.514	0.318	0.257	-0.061 ↓
2026 Nobel Chemistry laureate field	2.042	0.582	0.521	-0.061 ↓
Top MMLU score, year-end 2026	2.181	0.650	0.591	-0.060 ↓
2026 US CPI year-over-year, December print	1.623	0.254	0.311	+0.058 ↑
2026 US Senate majority	1.203	0.060	0.117	+0.057 ↑
2026 Wimbledon men's singles winner	0.938	0.086	0.031	-0.055 ↓
Top gold-medal nation, 2028 Summer Olympics	1.109	0.018	0.063	+0.045 ↑
Top SWE-Bench Verified score, year-end 2026	1.970	0.540	0.497	-0.043 ↓
2026 H1 SP500 return bucket	1.385	0.151	0.193	+0.041 ↑
2026 net global solar capacity addition	0.897	0.020	0.059	+0.039 ↑
2026 EU AI Act state of enforcement	1.816	0.447	0.471	+0.024 ↑
2028 US Presidential winner	1.050	0.009	0.029	+0.019 ↑
2026 NBA Finals winner conference	1.090	0.044	0.064	+0.019 ↑
2026 Pulitzer Prize Fiction winner setting	1.032	0.001	0.019	+0.018 ↑

(continued on next page)

(continued)

Partition	$\sum_t$	$\varepsilon_b^\star$	$\varepsilon_t^\star$	$\Delta\varepsilon^\star$
2026 H1 BTC return bucket	1.535	0.297	0.309	+0.012 ↑
2026 FIFA World Cup winner confederation	1.518	0.287	0.299	+0.012 ↑
Next FOMC decision (June 2026 meeting)	1.427	0.245	0.247	+0.002 ↑