Locally Coherent, Globally Incoherent: Detecting Compositional Incoherence in Multi-Component LLM Agents

Anany Kotawala [1]

Abstract

Multi-component LLM agents can be *locally coherent and globally incoherent*: each tool, subagent, or routed model call satisfies its own probability constraints, yet the composed output can violate joint consistency. We formalize this failure mode using the *compositional residual* $\varepsilon^\star$, the L_2 distance from the composed forecast to the joint coherent polytope. This quantity is a distribution-free certificate computable from the agent's output and the coupling constraints alone. A product-structure dichotomy characterizes when local coherence implies global coherence: under coordinate-selection aggregation, this holds if and only if the joint polytope factorizes as a Cartesian product of local polytopes. When coupling tightens this product, a hierarchical Boyle–Dykstra repair converges to the joint projection, yielding deterministic Brier improvement. The residual also supports online monitoring, repair, and stopping; an anytime-valid e-process over the residual stream provides uniform Type-I error control at any stopping time. Across $1{,}876$ ensemble cliques from four contemporary LLMs, $\varepsilon^\star$ is positive on 33%–94% of cliques by relation type; hierarchical JCD reduces it to numerical precision.

1. Introduction

Many contemporary LLM agent systems compose outputs from multiple modules. A planner may route retrieval to a search-augmented subagent, arithmetic to a calculator, and probability assessment to one or more forecasting calls, then pass the resulting probability claims to a downstream decision policy. We refer to such systems as *multi-component LLM agents*; the term includes multi-tool agents, specialist subagents, and routed ensembles. Each component contributes a subset of the joint belief on which the agent acts. Even when every component is calibrated and internally coherent on its assigned questions, the aggregated belief need not be coherent: a research component that emits $P(\text{Republican})=0.6$ and a forecasting component that emits $P(\text{Democrat})=0.6$ produce a 1.2-mass two-candidate quote that no probability measure can assign. This violates the probability axioms and induces a Dutch-book-style exposure (de Finetti, 1937) that arises strictly *between* components. We call this regime *locally coherent, globally incoherent*: per-component coherence repairs do not, in general, repair the composed system.

A worked example. Consider two components assigned complementary events. Component 1 quotes $P(A)=0.7$, which is locally coherent on $\{A\}$, and Component 2 quotes $P(\neg A)=0.6$, which is locally coherent on $\{\neg A\}$. The composed quote violates the negation constraint $P(A)+P(\neg A)=1$ by 0.3. The joint L_2 projection returns $P(A)=0.55$ and $P(\neg A)=0.45$; the *compositional residual* $\varepsilon^\star = \sqrt{0.0225+0.0225} \approx 0.21$ is the distance from the composed quote to the joint coherent polytope. Thus $\varepsilon^\star > 0$ certifies a Dutch-book exposure that is absent from either component in isolation (de Finetti, 1937).

This paper. We characterize when local coherence guarantees system coherence and when it does not, provide a constructive repair, and measure the resulting failure mode empirically. The compositional residual $\varepsilon^\star$, computable from the composed quote and coupling constraints alone, provides an online monitoring statistic, supports stopping and uncertainty communication, and extends to an anytime-valid sequential coherence test. Our contributions are:

- **Compositional guarantees.** A product-structure dichotomy (Theorem 3.3): under coordinate-selection aggregation, local coherence guarantees global coherence *for all inputs* if and only if the joint coherent polytope factorizes as a Cartesian product of local polytopes; whenever coupling makes the joint feasible set a strict subset of this product, there exist locally coherent forecasts whose composition is globally incoherent. The compositional residual $\varepsilon^\star$ certifies this failure mode (Corol-

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

Accepted at the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea.

lary 3.4), provides a deterministic norm-based bound on system-level Dutch-book exposure (Corollary 3.5), and is upper-bounded by inter-LLM disagreement (Proposition 3.7), making the existence claim quantitative.

- **Deterministic coherence control.** A constructive hierarchical repair via Boyle-Dykstra cyclic projection over local polytopes and the coupling set, converging in L_2 to the joint projection (Theorem 3.8). The repair lifts the per-call Pythagorean Brier improvement to the system level (Corollary 3.6); both guarantees are distribution-free in the deterministic sense.
- **Empirical evidence.** The unit-stake LMSR exposure diagnostic, a protocol-specific instantiation of the bound $\text{Exposure}^\star \leq \sqrt{m^\star}\,\varepsilon^\star$ from Corollary 3.5, averages \$0.13 per clique under naive composition and is reduced to the QP solver floor by the hierarchical JCD pipeline. The dollar figure is the LMSR-rescaled certificate, not realized financial loss; alternative elicitation rules rescale the constant but preserve the qualitative claim.
- **Sequential monitoring.** An anytime-valid coherence test (§3.6): the e-process E_t over the residual stream $(\varepsilon_t^\star)_{t\geq 1}$ is a supermartingale under the null of system coherence; the stopping rule $\tau_\alpha := \inf\{t : E_t \geq 1/\alpha\}$ controls Type-I error uniformly over all stopping times, giving an operator a statistically valid signal for escalating or halting an incoherent agent pipeline without committing to a fixed horizon (Theorem B.2).

2. Problem Setup

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; the standard projection facts below explain why this repair is also well behaved for proper scoring and local Dutch-book exposure.

Theorem 2.1 (Pythagorean Brier improvement)**.** *Let $p^* \in \mathcal{M}_C$ be the ground-truth marginal vector for a clique C and $\Pi(\hat{p})$ the L_2 projection of $\hat{p}$ onto $\mathcal{M}_C$. Then for every realization $\hat{p} \in \mathbb{R}^m$, $\|\Pi(\hat{p}) - p^*\|_2^2 \leq \|\hat{p} - p^*\|_2^2 - \|\hat{p} - \Pi(\hat{p})\|_2^2$. 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 $\langle \hat{p} - \Pi(\hat{p}),\, p^* - \Pi(\hat{p})\rangle \leq 0$. Expanding $\|\hat{p} - p^*\|^2$ around $\Pi(\hat{p})$ and dropping the non-negative cross term gives the result. □

Corollary 2.2 (Per-component finite-state FTAP)**.** *Treat $\Pi(\hat{p}) \in \mathcal{M}_C$ as quoted unit-pay prices on the m Bernoulli contracts of C. Then the finite-state Fundamental Theorem of Asset Pricing (FTAP) gives no state-independent Dutch book against the ideal projection. With a numerical QP solution, the unit-stake exposure diagnostic used below is bounded by $m \cdot \varepsilon^{\text{KKT}}$, where ε^{KKT} is the QP's primal residual on $\Pi(\hat{p})$ ($\leq 1.4\times 10^{-5}$ 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. Averaging and other linear or nonlinear aggregators are discussed in Section 5.

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. Compositional Residual and Repair

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). 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, $\Pi^\star(r) \neq r$ and $\varepsilon^\star = \|r - \Pi^\star(r)\|_2 > 0$. □

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 4 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 Hanson (2003)-LMSR 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 $\text{Exposure}^\star \leq \sqrt{m^\star}\varepsilon^\star$. 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 2.1 applied to $\Pi^\star$ on $\mathcal{M}^\star$. □

3.3. Disagreement controls the residual

Proposition 3.7 (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,*

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

A natural choice for r is the JCD-projected forecast $\Pi_\beta(\hat{p}^{(\beta)})$ of any component β that itself lies in $\mathcal{M}^\star$ (e.g., a component routed the entire joint clique). The right-hand side is the L_2 disagreement between the composed quote and any such reference; it vanishes when all coordinates are assigned to the same component (no composition; no disagreement). Inter-component disagreement on disputed coordinates is therefore a measurable proxy that helps explain and upper-bound 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 $x = \mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)})$. □

For the relations in our experiments, the bound takes a closed form. Writing $\Delta_j := \Pi_{\alpha(j)}(\hat{p}^{(\alpha(j))})_j - \Pi_\beta(\hat{p}^{(\beta)})_j$ for the per-coordinate disagreement between the assigned LLM $\alpha(j)$ and reference β:

- **Negation** ($p_1{+}p_2{=}1$, $a{=}(1,1)$): $\varepsilon^\star \leq |\Delta_1{+}\Delta_2|/\sqrt{2}$, with equality when only one constraint is active.
- **Partition** ($\sum_j p_j{=}1$, $a{=}\mathbf{1}_{m^\star}$): $\varepsilon^\star \leq |\sum_j \Delta_j|/\sqrt{m^\star}$.
- **Conjunction/disjunction** (Fréchet box, two half-space constraints): $\varepsilon^\star \leq \|\Delta\|_2$ by the 1-Lipschitz contraction of the projection.

Empirically the bound holds on 100% of the 1,876 ensemble cliques (verified post hoc by tightening each LLM's stored JCD output to OSQP exact-feasibility) and is informative: the mean bound-to-residual ratio is 1.30 on negations, 2.34 on disjunctions, 3.00 on conjunctions, and 4.39 on partitions. Cross-component disagreement is the dominant driver of compositional incoherence; the bound is tightest where the constraint normal a is most aligned with the disagreement direction (negation) and loosens for higher-dimensional polytopes (partition). Sequential applications of the same bound in the appendix audit (App. B) inherit the same tightness profile, so the per-relation ratios above transfer directly to the e-process construction.

Concrete instances of coupling. Three instances arise repeatedly in deployed multi-component agents and will recur in our experiments. (i) *Shared question disagreement:* two specialists both quote Q, but at different prices; the joint polytope demands they agree. (ii) *Negation across components:* sub-model a quotes Q, sub-model b quotes $\neg Q$; coupling constraint $p_a + p_b = 1$. (iii) *Partition across components:* the agent assembles a multi-candidate distribution from per-candidate probabilities elicited from different specialists; coupling constraint $\sum_a p_a = 1$.

3.4. 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 3.8 (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 and small-scale closed form. Each Π_a is the per-component JCD QP, $O(\text{OSQP})$ in m_a. Each $\Pi_\mathcal{C}$ is a projection onto an intersection of half-spaces in $[0,1]^{m^\star}$, also a small QP problem. In our experiments, the joint polytopes induced by the Paleka (neg/and/or) and Polymarket (partition) relation classes admit closed-form projections (see App. A and Wang & Carreira-Perpiñán (2013)); we use direct $\Pi^\star$ via the same QP machinery as the

per-call layer (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 tolerance reported. The QP path is also equivalent to converged Boyle-Dykstra over the same intersection and is substantially faster than running Dykstra to convergence on every clique.

3.5. Runtime use: monitor, repair, or escalate

The compositional residual is computable at runtime by any agent that observes its composed probability vector and the coupling constraints $\mathcal{C}$. This supports three deployment modes that share the same $\varepsilon^\star$ computation and differ only in how the agent acts on it.

- **Monitor-only.** The agent logs $\varepsilon^\star$ at each composition step and surfaces large residuals to an operator without altering its behavior. Useful when the policy maker wants visibility into compositional incoherence without intervening in the agent's quote.
- **Repair.** The agent replaces the naive composed quote $\mathcal{A}(\Pi_1, \ldots, \Pi_k)$ with $\Pi^\star(\mathcal{A}(\Pi_1, \ldots, \Pi_k))$ before passing it to downstream decision-making: the hierarchical JCD pipeline of Theorem 3.8. By Theorem 3.8 the repaired quote lies in $\mathcal{M}^\star$ up to solver tolerance, and by Corollary 3.5 the associated exposure diagnostic is reduced to numerical zero.
- **Abstain or escalate.** The agent refuses to act, or routes to a human, when $\varepsilon^\star > \tau$ for an operator-chosen threshold τ. The threshold is a risk–coverage knob; its calibration to budgeted exposure or to empirical conformal coverage is straightforward but outside our scope.

3.6. 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 B.1, App. B). 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 B.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 B.

4. Experiments

Setup. We evaluate on Paleka (Paleka et al., 2025) (negation, conjunction, disjunction, paraphrase relations; 134 cliques per checker) and Polymarket (67 partition cliques after leakage control, expanded to 268 ensemble instances by four seeds). Models: Anthropic `claude-haiku-4-5-20251001`, OpenAI `gpt-5.4-mini`, `gpt-5.4-nano`, Meta `llama-3.3-70b-versatile`, each producing $K{=}8$ verbalized 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 C. **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.

Planner-style routing simulation. To test whether the compositional residual appears outside synthetic benchmark ensembling, we constructed a controlled simulation replicating the information structure of a deployed multi-component agent: a planner LLM routes each candidate outcome to a randomly assigned specialist; each specialist returns a single Bernoulli marginal under a single-question prompt with no cross-candidate context; the planner assembles the resulting marginals into a partition quote. We applied this protocol to 100 live-style multi-candidate partition questions spanning politics, macroeconomics, sports, awards, technology, science, governance, health, and climate. *All* 100 composed quotes had positive compositional residual, with mean $\varepsilon^\star = 0.21$, median $\varepsilon^\star = 0.20$, and mean partition-mass violation $|\sum_i p_i - 1| = 0.39$. The largest violation was the "Largest US AI startup IPO of 2026" partition (four sectoral outcomes: AI infrastructure, model, applications, or other): the four assigned specialists each anchored to the assigned sector, yielding a partition mass of 2.50 and $\varepsilon^\star = 0.75$; ML-benchmark partitions (MMLU, SWE-Bench, HumanEval+, GPQA) had similarly large violations, each summing to ~ 2.0 as four specialists each anchored to the assigned candidate organization. This simulation indicates that the failure mode arises outside synthetic ensembling: standard specialist routing produces globally incoherent partition be-

liefs even when each specialist is locally coherent on its assigned Bernoulli question. Per-question quotes and assignment seeds for all 100 partitions appear in Appendix I.

4.1. Compositional residuals and the hierarchical fix

Sample sizes. For Paleka, N=536 per relation type (134 resolved cliques $\times$ 4 random-assignment seeds). For Polymarket partitions, N=268 (67 post-filter events $\times$ 4 seeds). Thus the random-assignment experiment has $536+536+536+268 = 1{,}876$ ensemble bets. Structured routing uses one deterministic assignment seed, giving $134+134+134+67 = 469$ bets.

The compositional residual is positive on most cliques. Figure 1a shows the empirical CDF of $\varepsilon_t^\star$ per relation type. The fraction of cliques with strictly positive $\varepsilon_t^\star$ is 94% on partition (Polymarket multi-candidate events), 66% on negation, 43% on disjunction, and 33% on conjunction. Mean $\varepsilon^\star$ ranges from 0.058 (and) to 0.118 (partition). *None* of this is visible at the per-component level, where each $\Pi_a(\hat{p}^{(a)}) \in \mathcal{M}_a$ has zero residual by construction.

The naive ensemble carries a non-negligible scale-normalized Dutch-book diagnostic that the hierarchical repair eliminates. Applying the unit-stake Dutch-book exposure diagnostic (Corollary 3.5; closed form per relation in Appendix A) to each ensemble clique, Figure 1b plots cumulative exposure for three regimes against a fixed clique ordering. The naive ensemble averages $\approx \$0.13$ per unit-stake clique on contracts paying at most \$1. We emphasize that this number is a *protocol-specific diagnostic*, not a universal agent-risk measure: it is what the Hanson-LMSR market-maker bound contracts to under unit stakes, and other elicitation protocols rescale the constant. Hierarchical JCD reduces exposure to 1.7×10^{-14}, which is the OSQP solver's numerical floor; Theorem 3.8 guarantees convergence of the iterates to the joint projection $\Pi^\star$, and this reported value is the empirical solver-tolerance gap to that projection (the theorem does not predict the constant). Note that this empirical floor is far below the worst-case per-clique bound $m^\star \cdot \varepsilon^{\mathrm{KKT}} \lesssim 10^{-4}$ from Corollary 2.2; that bound is a worst-case sum across the $m^\star$ unit-stake bets, and the empirical curve benefits from sign cancellation across them. The *no-composition reference* (a single LLM answers every coordinate; gray dashed) remains at the same QP floor by construction: each LLM is JCD-coherent on the full clique, so a single-LLM answer carries zero per-clique Dutch-book exposure deterministically. Thus hierarchical JCD removes the compositional exposure while retaining specialist routing, matching the no-composition reference at numerical floor.

Structured-routing topology. Random assignment is a stress test for cross-model mixing. A more structured alternative is *structured routing*, where each LLM is the specialist for one relation family (Llama-3.3-70b for atomic slots, Claude-Haiku-4.5 for negation slots, GPT-5.4-mini for conjunction/disjunction compounds, GPT-5.4-nano for partitions). Under structured routing the certificate is positive *more often* on negation (98.5% of cliques) and less often on the remaining relation families: conjunction 12%, disjunction 16%, and partition 36% (App. G). Random and structured routing bracket realistic deployments; the per-bet exposure is comparable (\$0.13 vs. \$0.11), but the relation breakdown depends sharply on assignment policy. Both regimes confirm the qualitative finding: cross-component incoherence can appear even when no single component has a local residual.

Worst-case examples. The five highest-$\varepsilon^\star$ cliques across both regimes appear in App. H: the worst is a negation pair where Claude quotes 0.84 for P and Llama quotes 0.89 for $\neg P$ (system mass 1.73, $\varepsilon^\star$=0.52); next come disjunctions where two specialists assign near-zero probability to the antecedents, while the disjunction specialist quotes $\sim$0.92 for $P \vee Q$.

Stratification by relation type. The four relation classes recover, at the system level, the same ordering by feasibility-polytope restrictiveness that the per-call analysis reveals at the component level (Table 2): partitions (94% of cliques exhibit cross-component incoherence) $>$ negation (66%) $>$ disjunction (43%) $>$ conjunction (33%). Partition is the most coupled relation ($m^\star-1$ degree-1 constraints); conjunction is the least (one Fréchet box). This ordering is consistent with the geometric intuition behind Theorem 3.3: tighter coupling enlarges the region of locally coherent but globally infeasible aggregates, although the typical residual magnitude depends on the distribution of inter-model disagreement.

4.2. Tool-augmented specialists

A natural concern for the no-tool experiment is that specialists may be quoting from incomplete training-time priors; a deployed multi-component agent may equip each component with retrieval. We test whether grounding each specialist with web search reduces the compositional residual. We re-run the first 30 partitions of the case study (same assignment seed) under a tool-augmented protocol: each specialist issues a single `web_search` query with its assigned outcome text, the top-five DuckDuckGo snippets are prepended to the prompt, and the specialist then emits its K=8 verbalized-probability samples conditional on this context. Compositional residual is computed identically; the only change is that each Bernoulli quote is now retrieval-grounded.

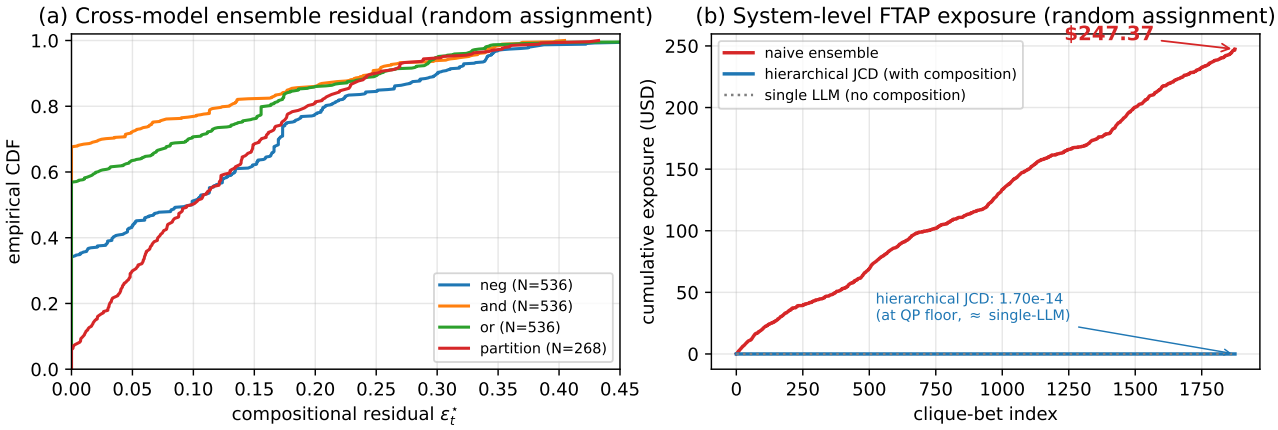

Figure 1. **Compositional residuals and the hierarchical fix.** *(a)* Empirical CDF of the compositional residual $\varepsilon_t^\star = \|\mathcal{A}(\Pi_1, \ldots, \Pi_k) - \Pi^\star(\,\cdot\,)\|_2$ across N=1,876 ensemble cliques, by relation type. Each component is individually JCD-projected so component residuals are zero by construction; the positive mass at $\varepsilon^\star > 0$ is *new* cross-component incoherence. *(b)* Cumulative unit-stake Dutch-book exposure diagnostic (Corollary 3.5) for three regimes: naive ensemble (red), hierarchical JCD (blue), and the no-composition reference (gray, dashed) in which a single LLM answers all coordinates of every clique. By JCD construction the no-composition reference is at the QP floor; hierarchical JCD recovers the same floor while *retaining* multi-component composition. The compositional gap is the entire $0.13/bet exposure of the naive curve.

Retrieval does not reduce $\varepsilon^\star$. Across the 30 matched partitions, mean $\varepsilon^\star$ shifts from 0.260 (no tool) to 0.283 (tool-augmented), with 20/30 partitions exhibiting equal or higher residual under retrieval (Figure 6, App. J). Per-specialist variance drops sharply (LLM responses tighten around the retrieved facts), but cross-specialist disagreement persists, because each specialist is grounding only on the snippets returned for *its own* outcome. The compositional failure is structural: each tool-using specialist is locally well-calibrated on its assigned Bernoulli, no specialist is responsible for the partition constraint, and the agent aggregates beliefs that no probability measure can simultaneously satisfy. Tool use sharpens individual claims but does not reconcile them across components; the JCD repair remains necessary.

4.3. Within-component checks

We next verify that the local JCD layer behaves as predicted by Theorem 2.1 before turning to the composition layer, which is the focus of this paper.

Brier improvement is conditional on incoherence. Within a single component, the slack $\|\hat{p} - \Pi(\hat{p})\|_2^2 = \varepsilon^2$ in Theorem 2.1 predicts that JCD's gain stratifies by pre-projection coherence violation. Sorting Paleka cliques by ε confirms the predicted ordering: the most-incoherent quartile gains 0.040 Brier on m=2 cliques and 0.013 on m=3, while the lower three quartiles see no statistically detectable improvement; on Polymarket the top-quartile gain reaches 0.12 at $m \geq 4$ (Figure 3 top panel). Per-question Platt calibration (B5 in Table 1) shows no such stratification, distinguishing the projection from generic calibration.

Among deployable marginal-repair baselines, B6 (Platt + JCD) achieves the lowest mean Brier on every (model, checker) cell. Component-level JCD reduces mean Brier on every (model, checker) cell of Table 5 (Paleka) and Table 4 (Polymarket); stacking with per-question Platt calibration (Platt, 1999) yields a 0.037-point Brier reduction on Paleka and 0.046 on Polymarket. Diebold-Mariano tests reject the null that JCD and the raw K-sample mean produce equal Brier with $p < 10^{-4}$ for every model (full significance table in App. E). A Murphy decomposition (Table 8) shows JCD reduces reliability error by 36% relative ($0.0357 \rightarrow 0.0229$) while *improving* resolution; calibration improvement, not over-extremization. B4 (direct joint elicitation) is an oracle-style per-call baseline that asks a single model for the full joint probability; it is *not* available in the modular setting of Section 3, where different components own different coordinates and the agent only observes composed marginal quotes. We include it to contextualize accuracy, not as a deployable replacement.

Per-relation gain. B6 reduces Brier by 21% on conjunctions, 37% on disjunctions, 4% on negations, 1% on paraphrase (Table 2); the ordering tracks the restrictiveness of the local polytope, predicted by Theorem 2.1. The within-component gain is small on negations because raw two-question marginals are already nearly coherent; the compositional layer in Section 4 is where negation-driven incoherence appears.

Table 1. Baselines and interpretation (per-component layer). B4 is an oracle that requires single-model batched joint elicitation and so is not available in the modular multi-component setting of Section 3; it appears here only to contextualize accuracy. B3 (prompted self-consistency) is a prompting-only baseline that increases Brier by $+43\%$ in the ablation; we retain it in the table because this negative result is informative: prompting and self-consistency methods do not substitute for geometric coherentization.

Baseline	**Tests**
B1: K=1 Karvetski projection	ordinary coherentization
B2: K-sample raw marginal mean	sampling alone
B3: prompted self-consistency	prompting alone
B4: direct joint elicitation	ask LLM for joint P
B5: per-question Platt calibration	calibration alone
B6: B5 + JCD	complementarity

Table 2. Per-relation B6 gain on Paleka, averaged across the four-model panel (N=536 (model, clique) evaluations per row).

Relation	B2	B6	Δ_{B6}
NegChecker	0.223	0.214	-4.2%
AndChecker	0.215	0.170	-20.9%
OrChecker	0.251	0.160	-36.5%
ParaphraseChecker	0.221	0.218	-1.2%

Reproducibility. Code, prompts, sample dumps, per-clique JCD residuals, the ensemble-assignment seed, and the full hierarchical-projection pipeline are specified in Section C. K-sweep and the sequential test appear in Appendix B.

5. Related Work and Limitations

On the convex-geometric foundations. The projection results used here, namely block-product factorization of L_2 projection, Hilbert projection theorem, Boyle–Dykstra cyclic projection, are textbook (Bauschke & Combettes, 2017; Boyle & Dykstra, 1986). Our contribution is not the geometry but its operational use: we identify the compositional residual $\varepsilon^\star$ as a runtime certificate computable by any multi-component agent that observes its composed quote, and we package the corresponding L_2 projection as a deployable hierarchical repair that retains specialist routing.

Position relative to coherentization. Prior coherentization (Predd et al., 2008; Wang et al., 2011; Karvetski et al., 2013) projects a panel of expert forecasts onto a coherent polytope as one-shot preprocessing before aggregation. Every expert forecasts the same clique, the projection sees every coordinate, and no constraints arise after the projection. A multi-component agent violates all three properties. Components own different coordinates: the agent assembles its quote from per-component marginals it routed for and never sees the underlying joint distributions. The coupling set $\mathcal{C}$, which captures shared questions, cross-component logical relations, and partitions spanning components, exists only *after* assembly, so a panel projection cannot represent it. The residual $\varepsilon^\star$ is a runtime quantity computable from the composed quote and $\mathcal{C}$ alone, without access to component internals.

Position relative to compositional uncertainty quantification. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021; Bates et al., 2021) and adaptive conformal inference (Gibbs & Candès, 2021) target per-output coverage; multi-step or multi-stage conformal guarantees are an active topic. Our guarantees are orthogonal: we control *cross-question* coherence within a multi-component agent, not per-question coverage. The two can be combined operationally: a conformal or calibration layer can be applied to each specialist's marginal forecast, after which $\varepsilon^\star$ can be computed on the composed joint quote to detect cross-question incoherence. If $\varepsilon^\star$ exceeds a threshold, the agent can repair or abstain as in §3.5. A formal joint coverage-coherence guarantee is outside our scope.

Agentic reliability and uncertainty communication. A growing body of work addresses uncertainty in single-model outputs, including calibration (Guo et al., 2017), self-consistency prompting (Wang et al., 2023), and conformal selective prediction, without addressing the inter-component coherence failure studied here; our ablation (B3) confirms that self-consistency increases Brier by $+43\%$ and does not repair cross-component coupling. Multi-step agent architectures that interleave tool calls and subagent outputs (Yao et al., 2023) inherit the same structural gap: each tool call may be individually well-formed, yet the composed belief can violate joint probability constraints. The e-process of §3.6 connects this certificate to safe sequential monitoring (Howard et al., 2021; Ramdas et al., 2023).

Limitations and extensions. Theorem 3.3 is existential; typical $\varepsilon^\star$ depends on inter-LLM disagreement, which we measure empirically but do not bound distributionally. Lemma B.1's envelope is loose for low-variance regimes and could be tightened by Bernstein refinements (Maurer & Pontil, 2009; Waudby-Smith & Ramdas, 2024). For non-linear aggregators the dichotomy need not hold, although $\varepsilon^\star$ and Boyle–Dykstra repair still apply on the joint polytope. Hierarchical JCD also requires a specified coupling set $\mathcal{C}$; extending the certificate to free-form planner-tool traces requires inferring those constraints.

References

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

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

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.

Gibbs, I. and Candès, E. J. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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.

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

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

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.

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

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.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

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

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

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.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Appendix: Supplementary Material

A. 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 $|\sum_i \hat{p}_i - 1|$.

B. Sequential test on the residual stream

In long-horizon agent deployments, the compositional residual stream $(\varepsilon_t^\star)_{t\ge1}$ 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 $(\varepsilon_t)_{t\ge1}$ that an agent generates over time admits a complementary *temporal* guarantee.

Setup. A clique stream $(C_t)_{t\ge1}$ 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 B.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 B.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 B.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 $\alpha \leq 10^{-4}$ within $\leq$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.

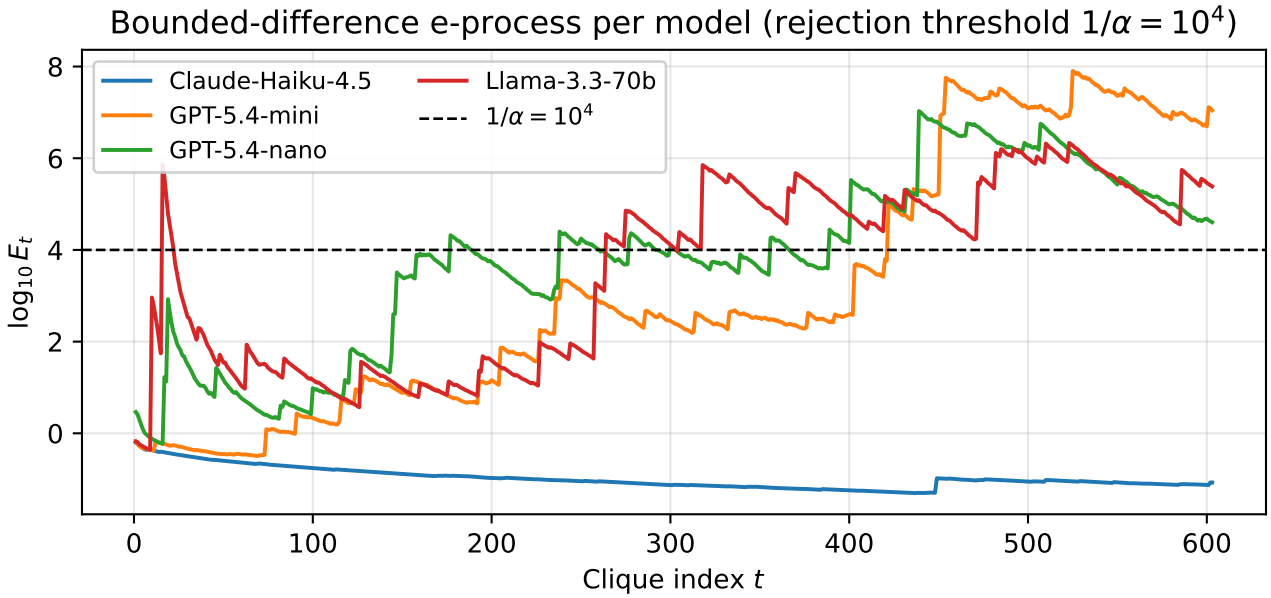

Figure 2. **Bounded-difference e-process E_t on per-component residual streams (Theorem B.2).** Crossing $1/\alpha = 10^4$ at t=16 (Llama), t=177 (GPT-nano), t=422 (GPT-mini); Claude-Haiku stays below 1.

C. Clique-mining pipeline and reproducibility

Paleka. Tuples are pre-mined by the original authors: each JSONL record is a single $(\text{checker}, \{Q_i\}, \text{relation})$ entry, 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 3.8 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 3 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 3. 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)

D. 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 4 gives the per-model breakdown.

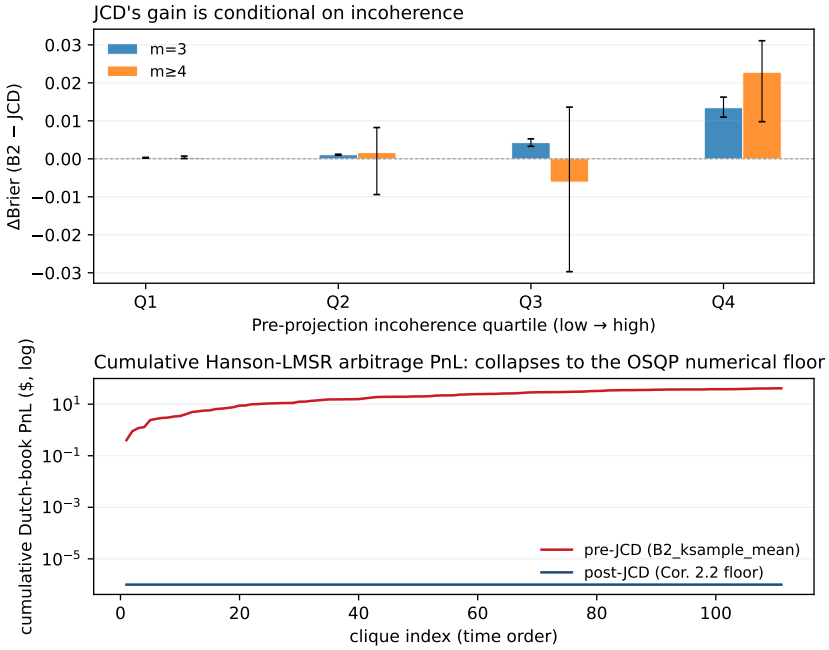

Figure 3. Polymarket cross-platform replication of the per-call result (Theorem 2.1): the incoherence-conditional gain pattern of Theorem 2.1 replicates on resolved Polymarket events, with the largest mean gain in the highest-ε quartile at $m \geq 4$ (note the smaller per-call absolute magnitudes; the lower panel's cumulative Hanson–LMSR exposure reduction to the QP floor matches Corollary 2.2).

Table 5. 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
azure_mini	0.2261	0.2112	0.2111	0.1886	0.1827	-0.0150	-0.0435
azure_nano	0.2356	0.2212	0.2209	0.2015	0.1989	-0.0148	-0.0367
groq_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

Table 6. 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 4. 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

E. Full per-cell tables

Per-model Brier under each baseline (Table 5), the six-method ablation including B3/B4 (Table 6), the significance tests (Table 7), and the Murphy reliability/resolution decomposition (Table 8).

F. K-sweep on real LLM forecasts

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.

G. 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.

Table 7. 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
azure_mini	10.682	1.657e-24	0.0150	0
azure_nano	12.281	4.277e-31	0.0148	0
groq_llama-3.3-70b-versatile	7.980	7.437e-15	0.0079	0

Table 8. 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

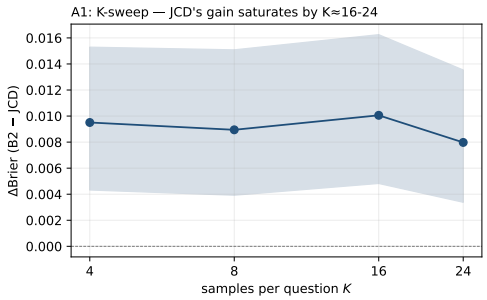

Figure 4. K-sweep on Anthropic Claude-Haiku-4.5 NegChecker forecasts. Non-vanishing ΔBrier asymptote: the empirical signature of structural incoherence in p_F.

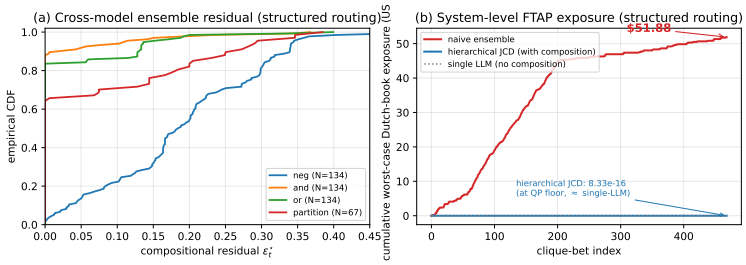

Figure 5. **Structured-routing replication of Figure 1.** The same certificate is positive with a different relation breakdown but comparable per-bet exposure (\$0.11/bet vs. \$0.13/bet under random assignment). The lower cumulative endpoint reflects fewer bets under structured routing, N=469 vs. N=1,876,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

H. Worst-case failure modes

Table 9 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 9. 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

I. 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 10. 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

(continued on next page)

(continued from previous page)

Partition	m	per-outcome quote (LLM)	$\sum p_i$	$\varepsilon^\star$
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
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

(continued on next page)

(continued from previous page)

Partition	m	per-outcome quote (LLM)	$\sum p_i$	$\varepsilon^\star$
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
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.

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

We re-run the first 30 partitions of the case study under a tool-augmented protocol (§4.2). 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.

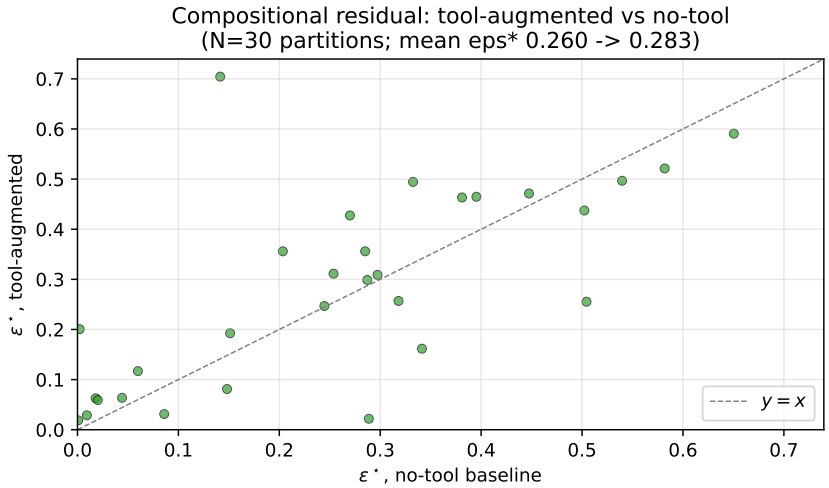

Figure 6. **Tool augmentation does not eliminate the compositional residual.** Each point is one of the 30 matched partitions; x-axis is $\varepsilon^\star$ in the no-tool baseline, y-axis is $\varepsilon^\star$ when each specialist receives top-five DuckDuckGo snippets for its assigned outcome before 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 11. 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 ↑

(continued on next page)

(continued)

Partition	$\sum_t$	$\varepsilon_b^\star$	$\varepsilon_t^\star$	$\Delta\varepsilon^\star$
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 ↑
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 ↑