Locally Coherent, Globally Incoherent: A Post-Aggregation Failure Mode in Multi-Component LLM Agents

Anany Kotawala [1]

Abstract

Multi-component LLM agents (planners with tools, specialist routers, multi-stage pipelines) assemble probabilistic claims from components that each see only part of a joint problem. Each component can be locally coherent, and even individually calibrated in principle, yet the composed belief can violate basic probability axioms: marginals over complementary events sum past one; partitions exceed the unit-mass budget by more than $2\times$. Three intuitive mitigations (retrieval grounding, partition-aware prompting, and delegating coherentisation to an aggregator LLM) each silently fail or actively regress on a measurable fraction. We trace the failure to specialists not seeing cross-component coupling constraints, formalise it as an output-only runtime certificate, and propose a deterministic geometric repair that drives the residual to zero without an LLM call. The failure is structural: $\varepsilon^\star > 0$ on 33–94% of 1,876 ensemble cliques, $20/20$ planner-discretion partitions, and 97.8% of 268 frontier-panel bets, translating to $+0.115$ nats per bet of decision-relevant regret on 1,770 resolved markets. Disclosing the constraint to specialists reduces but does not eliminate the residual. Code: https://github.com/ananykotawala/composition-incoherence-icml.

1. Introduction

Modern LLM agents decompose probabilistic reasoning across components: a planner routes subqueries to specialist sub-agents or tools (Yao et al., 2023; Wei et al., 2022), each specialist handles one slice of the joint problem in isolation, and an aggregator assembles their outputs into a joint belief (Cemri et al., 2025; Wornow et al., 2025). Components communicate through outputs alone (prompts, JSON fields, function returns), so cross-component coupling constraints – shared questions, partitions, cross-component logical relations – are invisible at elicitation time and only materialise after assembly. The composed belief can therefore violate the basic probability axioms (de Finetti, 1937) even when every component is locally coherent on its own slice. Figure 1 shows a four-specialist instance from our planner-discretion routing harness (§5.2): each specialist anchors a single sector of a partition at a probability below one, but the four contributions sum to 2.50 – the assembled belief assigns two-and-a-half times the unit mass admissible under the partition constraint $\sum_i p_i = 1$. The incoherence is invisible in any single specialist trace; it appears only after the agent composes the outputs. We call this failure mode *locally coherent, globally incoherent*: a post-aggregation failure of agentic AI.

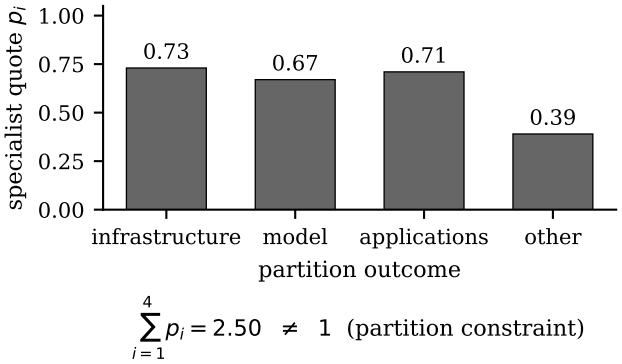

Figure 1. Per-sector specialist quotes on the "Largest US AI startup IPO 2026" partition from the planner-discretion routing harness (§5.2). Each specialist is routed to one outcome and quotes a probability p_i; the four quotes sum to 2.50, violating the partition constraint $\sum_i p_i = 1$. No specialist saw the partition at elicitation time.

Existing tools do not cover this regime. Per-component calibration (Guo et al., 2017; Wang et al., 2023) operates on each slice in isolation; classical expert-aggregation (Predd et al., 2008; Wang et al., 2011; Karvetski et al., 2013) assumes every expert forecasts every coordinate, and so does not see the cross-component coupling constraints (shared questions, partitions, logical relations) that only arise after assembly. Multi-agent failure-mode taxonomies (Cemri

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

Accepted at the ICML 2026 Workshop on Failure Modes of Agentic AI (FAGEN), Seoul, South Korea.

et al., 2025; Zhu et al., 2025; Schmied et al., 2025) annotate qualitative categories from execution traces, but provide no quantitative runtime certificate and no repair that runs without further LLM calls.

We propose a *compositional residual* $\varepsilon^\star$, defined as the L_2 distance from the composed quote to the joint coherent polytope, that an operator can compute from the agent's output and the cross-component coupling constraints alone – without access to component internals. $\varepsilon^\star$ vanishes exactly when local coherence forces joint coherence and otherwise certifies a Dutch-book exposure on the assembled belief. We then exhibit a deterministic geometric repair – a hierarchical Boyle–Dykstra projection (Boyle & Dykstra, 1986; Bauschke & Combettes, 2017) – that drives $\varepsilon^\star$ to zero in milliseconds without an extra LLM call. A Brier improvement of -0.018 corresponds to roughly $+0.115$ nats of decision-relevant information per partition query under proportional allocation (§5.7).

Contributions. (i) The compositional residual $\varepsilon^\star$ as an output-only runtime certificate of cross-component incoherence; (ii) a product-structure dichotomy (Theorem 4.2) characterising when local coherence guarantees global coherence; (iii) a deterministic hierarchical Boyle–Dykstra repair that drives $\varepsilon^\star$ to the QP numerical floor with no extra LLM call; (iv) negative results on three intuitive LLM-side mitigations (retrieval, partition-aware prompting, and LLM-as-aggregator), each of which fails to eliminate or actively regresses $\varepsilon^\star$ on a measurable fraction.

Sections 3 and 4 formalise $\varepsilon^\star$, prove the dichotomy, and give the geometric repair. Section 5 measures the failure under planner routing at mid-tier and frontier scale, evaluates three LLM-side mitigations (retrieval, partition-aware prompting, LLM-as-aggregator), and tests a paired coupling-visibility mechanism probe. Section 6 relates the contribution to existing failure-mode taxonomies.

2. Related Work

Coherentisation and uncertainty quantification. Prior coherentisation (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: much classical expert-aggregation focuses on panel-level forecast repair, often before deployment-time agent assembly, and the post-aggregation coupling constraints induced by component ownership are not the primary object of study. Block-product L_2 factorisation, Hilbert projection, and Boyle–Dykstra cyclic projection are textbook (Bauschke & Combettes, 2017; Boyle & Dykstra, 1986); the contribution here is operational. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021; Gibbs & Candès, 2021) targets per-output coverage on an axis orthogonal to cross-question coherence; the two compose. The analogous KL/I-projection variant gives a Pythagorean log-loss improvement and is the natural choice when the downstream consumer is log-likelihood-based; we use L_2 for the closed-form partition-simplex projection and the Brier-side analysis.

Failure modes in agentic AI. Recent workshops characterise multi-agent failure taxonomies (Cemri et al., 2025), agent error attribution (Zhu et al., 2025), decision-making failure modes (Schmied et al., 2025), and trustworthiness benchmarks (Wornow et al., 2025). Self-consistency prompting (Wang et al., 2023) does not repair cross-component coupling (B3: $+43\%$ Brier). Multi-step architectures (Yao et al., 2023) inherit the structural gap; the e-process construction connects $\varepsilon^\star$ to safe sequential monitoring (Howard et al., 2021; Ramdas et al., 2023).

3. 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. 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 (de Finetti, 1937). 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 within-component projection inherits the standard guarantees (Pythagorean Brier improvement Theorem B.1 and a finite-state FTAP no-Dutch-book bound Corollary B.2), stated and proved in Section B.

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. We scope the dichotomy (Theorem 4.2) to the *owner-selected* coordinate selector of Definition B.4: each joint coordinate is the unique output of one component. This covers the deployment patterns evaluated here – planner-to-specialist routing, disjoint-tool composition, and sharded retrieval pipelines. Aggregators that mix outputs across components (averaging ensembles, logarithmic opinion pools, debate or voting protocols, mixture-of-experts) lie outside the iff. The certificate $\varepsilon^\star$ and the Boyle–Dykstra repair still apply on the joint polytope under any such aggregator; only the existential direction of Theorem 4.2 requires the owner-selected hypothesis. The averaging case is discussed in Section 6.

With $\hat{p} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})$ and aggregator $\mathcal{A}$ fixed, the local-then-global workflow $\Pi^\star(\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)}))$ and the direct-global workflow $\Pi^\star(\mathcal{A}(\hat{p}^{(1)}, \ldots, \hat{p}^{(k)}))$ both project into $\mathcal{M}^\star$; the substantive question is when, and by how much, they differ.

4. Method

4.1. Definitions

Definition 4.1 (Compositional residual)**.** For inputs $\hat{p} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(k)})$ and aggregator $\mathcal{A}$, let $q := \mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)})$ denote the composed quote; then

$$\varepsilon^\star(\hat{p}) := \left\| q - \Pi^\star(q) \right\|_2.$$

For each component a lift $\mathcal{M}_a \subseteq [0,1]^{m_a}$ to $\mathcal{M}_a^\uparrow \subseteq [0,1]^{m^\star}$ (free on non-component coordinates) and write $\mathcal{M}^\boxtimes := \bigcap_a \mathcal{M}_a^\uparrow$ for the joint feasibility under local constraints alone; $\mathcal{M}^\star \subseteq \mathcal{M}^\boxtimes$ with equality iff $\mathcal{C}$ is trivial. Owner-selected aggregation (Definition B.4, App. B) assigns each joint coordinate to a single component; the framework subsumes averaging on disjoint coordinates and on shared-coordinate equality couplings (Remark B.5, App. B).

4.2. Mechanism: why local coherence is not preserved

Why does the failure happen? Each specialist sees only its own local coordinates and the constraints over them. The agent acts on a joint object whose feasible set respects coupling constraints (shared questions, partitions, cross-component logical relations) that no single specialist can observe. Coupling *tightens* the joint feasible set below the Cartesian product of local feasible sets; whenever this tightening is strict, there exist locally coherent specialist forecasts whose composition lies outside the joint polytope. Theorem 4.2 formalises this as an iff.

Theorem 4.2 (Product-structure dichotomy)**.** *Under owner-selected coordinate aggregation, $\varepsilon^\star \equiv 0$ for all $\hat{p}$ iff $\mathcal{M}^\star = \mathcal{M}^\boxtimes$. When $\mathcal{M}^\star \subsetneq \mathcal{M}^\boxtimes$, there exist locally coherent inputs whose composition has $\varepsilon^\star > 0$. Equivalently, on a given input, $\varepsilon^\star(\hat{p}) = 0$ iff $\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)}) \in \mathcal{M}^\star$.*

(Forward and reverse-existential proofs in Section B.)

Exposure and Brier improvement (statements only). If $\varepsilon^\star > 0$, de Finetti's finite-state FTAP yields a Dutch-book portfolio with $\mathrm{Exposure}^\star \leq \sqrt{m^\star}\,\varepsilon^\star$ (Corollary B.6). For any composed quote r and joint truth $p^\star \in \mathcal{M}^\star$, $\|\Pi^\star(r) - p^\star\|_2^2 \leq \|r - p^\star\|_2^2 - (\varepsilon^\star)^2$ (Pythagorean Brier improvement, Corollary B.7). Both proofs (Cauchy–Schwarz on the defining halfspace; Hilbert projection theorem) are deferred to Section B.

4.3. Disagreement controls the residual

Proposition 4.3 (Disagreement upper bound)**.** *For any reference $r \in \mathcal{M}^\star$, $\varepsilon^\star(\hat{p}) \leq \|\mathcal{A}(\Pi_1\hat{p}^{(1)}, \ldots, \Pi_k\hat{p}^{(k)}) - r\|_2$.*

The bound is the L_2 disagreement between the composed quote and any coherent reference (proof in Section B, immediate from the variational characterisation of L_2 projection). Closed forms per relation in Section J; empirically the bound holds on 100% of 1,876 ensemble cliques and is strongly *predictive* of the residual: Spearman $\rho(\text{bound}, \varepsilon^\star)$ ranges from 0.83 (partition) to 0.97 (negation) on positive-residual cliques.

4.4. Hierarchical Boyle–Dykstra repair

The repair is the cyclic L_2 projection over the family $\{\Pi_1, \ldots, \Pi_k, \Pi_\mathcal{C}\}$ onto $\{\mathcal{M}_a^\uparrow\}$ and $\mathcal{C}$, whose intersection is $\mathcal{M}^\star$. The Boyle–Dykstra iteration converges to $\Pi^\star(r_0)$ from any starting point (Boyle & Dykstra, 1986; Bauschke & Combettes, 2017); statement and proof sketch in Theorem B.3, Section B.

What this fix costs. *Latency:* one OSQP projection per composition step ($\leq$ a few ms in $m^\star \leq 8$). *Capability:* none – the repair preserves all per-component information and only redistributes mass to satisfy coupling constraints. *Generalisation:* requires an explicit coupling set $\mathcal{C}$. For relation-typed cliques (negation, partition, conjunction, disjunction) $\mathcal{C}$ is given; for free-form planner-tool traces $\mathcal{C}$ must be inferred from the planner's prompt structure, which we leave to future work. On the relations evaluated, the repair admits closed-form one-line projections for partition and negation and requires the cyclic projection on conjunction and disjunction (Section I).

$\varepsilon^\star$ supports anytime-valid sequential monitoring via an e-process; full construction in Section D.

5. Results

5.1. Experimental setup

Datasets. 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. Claude-Haiku-4.5, GPT-5.4-mini, GPT-5.4-nano, and Llama-3.3-70b-versatile, each producing $K{=}8$ verbalized samples per question at temperature 0.7. Exact model identifiers and infrastructure are listed in the reproducibility manifest (Section E).

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 the source benchmark (Paleka et al., 2025). Full filtering protocol in Section 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.

5.2. Planner-discretion routing harness

We ask whether the failure mode persists when a planner LLM chooses both routing and per-specialist prompts with no hard-coded protocol. This rules out the failure being a routing-policy or prompting artefact.

For each of 20 partitions a planner LLM (Claude-Haiku-4.5) sees the partition label and full list of outcomes, picks a specialist per outcome from the four-LLM roster, and emits a per-specialist context block it judges sufficient for calibrated quoting (it may include the partition structure or any other context). Specialists sample $K{=}8$ verbalized probabilities under the planner's prompt; we assemble the marginals and compute $\varepsilon^\star$ as in §5.1.

Mean $\varepsilon^\star = 0.113$, versus 0.231 on the same 20 partitions under random-assignment routing (§5.1): a 49% reduction; $15/20$ partitions improve. Of the $5/20$ that worsen, four had baseline $\varepsilon^\star < 0.05$ and the fifth (US Senate majority, $0.060 \to 0.202$) is a counterexample where the planner's chosen routing introduced incoherence on a moderately-coherent baseline. Crucially, $\varepsilon^\star > 0$ on $\mathbf{20/20}$ partitions: the compositional failure mode persists when the planner has full discretion. The planner's routing histogram (GPT-mini 50%, Claude-Haiku 31%, Llama-70b 13%, GPT-nano 6%) shows no self-routing to its own family, ruling out that confound.

5.3. Mechanism probe: coupling-visible elicitation

The mechanism claim of §4.2 is that $\varepsilon^\star$ arises because specialists do not see the coupling constraint. We test this directly. On the same 20 partitions of §5.2, we re-elicit each specialist's quote under two conditions: BLIND (same planner-chosen routing as §5.2, but the planner-emitted context is removed and the specialist sees only its assigned outcome as a single Bernoulli question) and INFORMED (same routing; additionally the specialist sees the partition label, all sibling outcomes, the explicit $\sum_i p_i = 1$ constraint, and the peers' BLIND quotes on the other outcomes). Same four specialists, $K{=}8$, $T{=}0.7$. Each (partition, condition) cell is replicated across four independent $K = 8$ sampling rounds ("seeds"), paired BLIND→INFORMED within each round; the planner-harness has no random-assignment seeds, so the four seeds here vary sampling stochasticity, not routing. Note that INFORMED varies two things at once – coupling visibility *and* peer-quote disclosure – so the comparison identifies *coupling-visible elicitation with peer context* as the intervention, not constraint visibility in isolation. Full setup in Section M.

The intervention reduces $\varepsilon^\star$ as predicted: paired $\Delta\varepsilon^\star = \varepsilon^\star_{\text{BLIND}} - \varepsilon^\star_{\text{INFORMED}} = \mathbf{+0.221}$ (95% paired-bootstrap CI $[+0.173, +0.270]$; Wilcoxon $W{=}257$, $p = 2.6 \times 10^{-10}$; $n_{\text{pairs}} = 77/80$ cells surviving parsing across 20 partitions $\times$ 4 seeds, 3 lost to parse failures), with $16/20$ partitions improving on the per-partition mean. Nine of the 20 partitions substituted Llama-4-Maverick for Llama-3.3-70b due to a host-side rate limit during BLIND elicitation; each partition is internally consistent (one Llama variant throughout); details in Section M. But $\varepsilon^\star$ does not collapse to zero – mean $\varepsilon^\star_{\text{INFORMED}} = 0.298$ versus $\varepsilon^\star_{\text{BLIND}} = 0.519$, a 43% reduction. (Note: BLIND mean 0.519 exceeds the random-assignment baseline of 0.231 used in §5.2 on the same 20 partitions because BLIND strips even the planner's emitted context block, isolating the pure single-Bernoulli baseline; §5.2 retains it.) Specialists respond substantially to the disclosed peer quotes (mean $|\Delta q| = 0.25$ per outcome; 77% of outcomes shift by > 0.05) but cannot reconcile coordinate-by-coordinate. The mechanism prediction is supported; the geometric repair is still required to drive $\varepsilon^\star$ to the QP floor.

5.4. Random-assignment ensemble at scale

The planner-discretion harness (§5.2) established the failure mode on a real agent harness; we now show it persists under random-assignment ensembles at scale.

Rigid-routing scaling simulation (100 partitions). To verify the failure mode at scale outside synthetic benchmark ensembling, we ran a controlled simulation on 100 live-style multi-candidate partition questions (politics, macro, sports, awards, tech, science, governance, health, climate): a planner LLM routes each outcome to a randomly assigned specialist that returns a single Bernoulli under a single-question prompt; the planner assembles a partition quote. *All* 100 composed quotes had positive $\varepsilon^\star$ (mean 0.21, mean partition-mass violation 0.39); per-question detail in Section H.

Random-assignment ensemble. $N{=}3 \times 536 + 268 = 1{,}876$ ensemble bets (536 per non-paraphrase Paleka relation, 268 Polymarket). The fraction of cliques with $\varepsilon_t^\star > 0$ is 94% on partition, 66% on negation, 43% on disjunction, 33% on conjunction (Figure 2a). Mean $\varepsilon^\star$ ranges from 0.058 (and) to 0.118 (partition). *None* of this is visible per-component, where $\Pi_a(\hat{p}^{(a)}) \in \mathcal{M}_a$ has zero residual by construction.

Dutch-book exposure diagnostic. The unit-stake exposure diagnostic (Corollary B.6, closed form per relation in Section C) averages $\approx$\$0.13/bet under naive composition; hierarchical JCD reduces this to the OSQP numerical floor 1.7×10^{-14}, matching the no-composition reference (Figure 2b). This is a protocol-specific diagnostic; alternative elicitation rules rescale the constant.

Robustness checks. Under *structured routing* (each LLM owns one relation family), per-bet exposure is comparable (\$0.11 vs. \$0.13) with a different relation breakdown (Section F). Top-5 worst cliques are in Section G; the worst is a negation pair ($P{=}0.84$, $\neg P{=}0.89$, $\varepsilon^\star{=}0.52$). The failure mode is *structural, not sampling noise*: a K-sweep across $K \in \{4, 8, 16\}$ holds mean $\varepsilon^\star \approx 0.20$ flat on the Claude-Haiku NegChecker cell (Section L). The downstream-regret result of §5.7 is computed against ground-truth resolutions and is K-independent by construction.

Per-specialist residual attribution. Per-clique we attribute residual mass to each specialist by the L_2 norm of $(x - \Pi^\star(x))$ restricted to the coordinates that specialist owned (Section A). Across the 894 positive-residual bets, mean attributed L_2 mass is Claude-Haiku 0.092 ($n{=}506$ cliques where it owned at least one coordinate), GPT-5.4-mini 0.120 ($n{=}605$), GPT-5.4-nano 0.093 ($n{=}491$), Llama-3.3-70b 0.098 ($n{=}464$). GPT-mini carries the largest share among the four mid-tier specialists; the frontier-panel rerun (§5.6) tests whether capability scaling shifts this attribution.

5.5. LLM-side mitigations

We test three obvious LLM-side mitigations against the geometric repair: retrieval grounding, partition-aware prompting, and delegating coherentisation to an aggregator LLM. Table 1 summarises; details follow.

Outcome-local retrieval grounding. Each specialist issues one DuckDuckGo `web_search` on its assigned outcome text; the top-five snippets are prepended to the prompt and $K{=}8$ samples are drawn conditional on this context. Mean $\varepsilon^\star$ shifts from 0.260 (no tool) to 0.283 (tool-augmented), with 20/30 partitions equal or worse under retrieval (Figure 7, Section R). Per-specialist variance drops sharply but cross-specialist disagreement persists because each specialist grounds only on snippets for its own outcome.

Partition-aware prompting. Each specialist instead receives a prompt listing all sibling outcomes and stating its quote will be combined into a partition. Mean $\varepsilon^\star$ drops $0.260 \to 0.066$ (-75%); 25/30 partitions improve. The remaining 5/30 all *worsen*, and they are precisely those whose baseline residual was already below 0.05 (e.g. "2028 US Presidential winner" $0.009 \to 0.047$, "2026 Pulitzer Prize Fiction" $0.001 \to 0.030$): partition-aware framing nudges already-coherent specialists toward a generic mean and introduces incoherence the bare-prompt baseline did not have.

LLM-as-aggregator delegation. A natural alternative is to delegate coherentisation to an aggregator LLM. We hand the assembled marginals, outcome texts and partition label to Claude-Haiku-4.5 at $T{=}0$ and ask for a coherent partition vector (Section Q). Across the 100 partitions, the LLM aggregator drops mean $\varepsilon^\star$ from 0.214 to 0.028 – but coherence ($\varepsilon^\star < 10^{-3}$) is achieved on only 81/100, $\varepsilon^\star < 10^{-2}$ on 83/100, and the aggregator *actively regresses on* 7/100 (e.g. "2026 Nobel Physics laureate field": naive $0.318 \to$ LLM 0.500 with output sum 2.000, in direct violation of the explicit "sum to 1" instruction). On 0/100 partitions does the LLM aggregator beat hierarchical JCD; the geometric repair achieves $\varepsilon^\star \leq 1.5 \times 10^{-16}$ on all 100. The LLM aggregator's failures are silent; $\varepsilon^\star$ on the assembled quote is a verifiable certificate.

Gating implication for LLM-side mitigations. Combined with the planner-discretion harness (§5.2), 9/10 worsening cases across both prompt-side interventions had baseline $\varepsilon^\star < 0.05$ vs. a marginal rate of $\sim 20\%$. Prompt-only and LLM-aggregator mitigations can be harmful on already-coherent quotes; the certificate gates this, applying repair (geometric or prompt-based) only when $\varepsilon^\star > \tau$.

5.6. Frontier-model evaluation

We rerun the random-assignment ensemble on the 67 Polymarket partition cliques with a frontier-only specialist roster: Claude-Opus-4.7, GPT-5.5 (with reasoning disabled for apples-to-apples), DeepSeek-V3.2, and Llama-4-Maverick-17B-128E (FP8). The protocol matches §5.1 exactly ($K{=}8$, simplex pre-projection, 4 seeds, 268 bets; Section N).

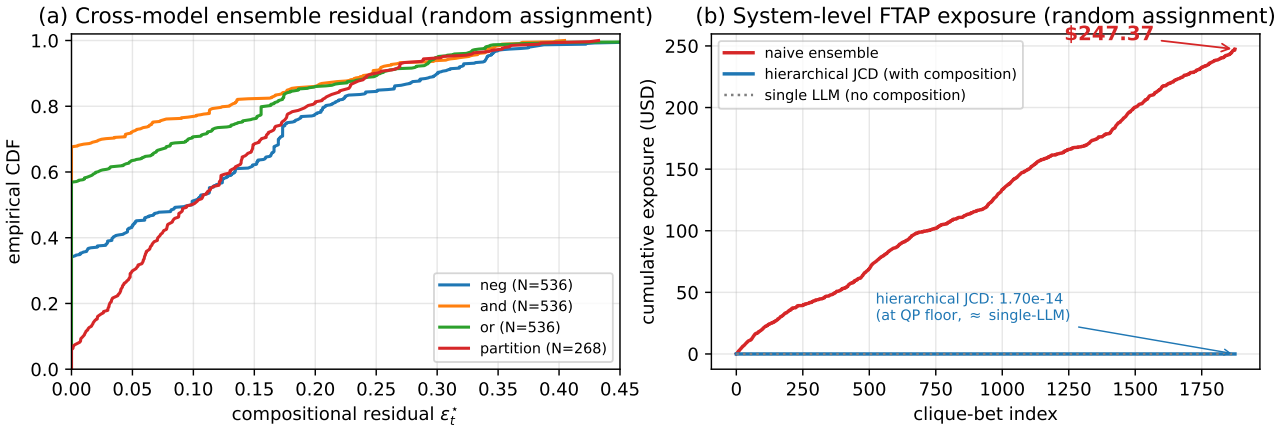

Figure 2. **Compositional residuals and the hierarchical fix.** *(a)* Empirical CDF of $\varepsilon^\star$ across N=1,876 ensemble cliques, by relation type. Each component is individually JCD-projected so component residuals are zero by construction; the positive mass at $\varepsilon^\star>0$ is *new* cross-component incoherence. *(b)* Cumulative unit-stake Dutch-book exposure (Corollary B.6) for three regimes: naive ensemble (red), hierarchical JCD (blue), and the no-composition reference (gray, dashed) in which one LLM answers all coordinates. Hierarchical JCD recovers the no-composition floor while retaining multi-component routing.

Table 1. The three LLM-side mitigations (retrieval-augmented, partition-aware prompt, LLM-as-aggregator) reduce $\varepsilon^\star$ in some settings but either fail to eliminate it or regress on a measurable fraction; the geometric repair is deterministic and requires no extra LLM calls. Rows use the evaluation subset for the corresponding mitigation: retrieval and partition-aware prompt on 30 partitions; naive, LLM-aggregator, and JCD on 100 partitions.

Method	mean $\varepsilon^\star$	$\Pr(\varepsilon^\star>0.05)$	% regress vs. naive	cost/query
Naive composition	0.214	0.86 (86/100)	–	0 extra calls
Retrieval-augmented	0.283	0.87 (26/30)	20/30	1 search/specialist
Partition-aware prompt	0.066	0.53 (16/30)	5/30	0 extra calls
LLM-as-aggregator	0.028	0.15 (15/100)	7/100	1 LLM call
Hierarchical JCD (ours)	$\leq 1.5\times10^{-16}$	0	0	1 QP solve (∼ms)

$\varepsilon^\star > 0$ on **97.8**% of the 268 clique-seed bets (67 unique partition cliques $\times$ 4 random-assignment seeds; the bootstrap is paired by clique-seed) – slightly *higher* than the mid-tier panel's 94%. Mean $\varepsilon^\star$ drops from 0.118 to 0.072 (-39%): on this partition slice, frontier models reduce residual magnitude but do not eliminate the failure. Brier improvement remains highly significant (-0.012, 95% CI $[-0.015, -0.009]$); log-payoff trends positive ($+0.107$ nats per bet) but is noisier on this slice.

5.7. Downstream-decision regret

The compositional residual is a geometric distance; we now show it translates into measurable decision harm. For each of 1,770 ensemble bets retained by the random-assignment protocol of §5 (we drop 106/1,876 bets where an assigned specialist lacks a forecast for the assigned coordinate, a data-pipeline artefact on partition events; see Section K), we converted the composed quote $\hat{p}^\star$ into a proportional allocation $w_i = \max(\hat{p}_i^\star, 0)/\sum_j \max(\hat{p}_j^\star, 0)$ over the clique outcomes and measured (i) the Brier loss against the binary resolution vector and (ii) the realised log-payoff $\log w_{\text{winner}}$ on the 1,250 bets where exactly one coordinate resolved YES. The proportional rule is *favourable to naive composition* – it rescales away partition-mass excess – so the JCD advantage we report below is a lower bound on what stricter rules (e.g. truncated Kelly without re-normalisation) would show.

Hierarchical JCD reduces mean Brier by -0.018 (95% paired-bootstrap CI over (clique, seed) pairs: $[-0.021, -0.015]$) and improves mean log-payoff by $+0.115$ nats per bet ($[+0.056, +0.187]$); both effects are highly significant. JCD recovers single-LLM-oracle Brier (oracle: $\Delta\text{Brier} = -0.017$, CI $[-0.035, -0.000]$) but with a confidence interval $\sim 8\times$ narrower: the projection is *reliably* above naive whereas the oracle's gain is high-variance across the four-LLM panel. The largest per-relation effect is on partition cliques, where naive composition loses 0.60 nats per bet vs. JCD ($[+0.15, +1.15]$, n=162) – roughly 40% of a bit of decision-relevant information per query. Brier (a proper score) is allocation-rule-independent; log-payoff is rule-dependent and shrinks to $+0.006$ nats under bettor strategies that themselves coherentise the quote (truncated Kelly, max-entropy; Section P).

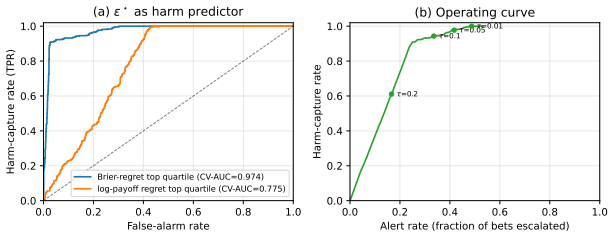

Figure 3. $\varepsilon^\star$ **as a runtime gate.** *(a)* ROC of $\varepsilon^\star$ predicting top-quartile realised regret on the 1,770-bet ensemble. Log-payoff curve (orange) is independent of the geometric residual; Brier curve (blue) is geometrically tied to $\varepsilon^\star$ by Pythagoras. *(b)* Operating curve: fraction of harmful bets caught vs. fraction of bets escalated, with τ values annotated. The high-recall point $\tau = 0.15$ catches 91% of harm at 25% alert rate.

5.8. Runtime gating thresholds

The deployment recommendation is to apply repair (or escalate) only when $\varepsilon^\star > \tau$. We calibrate τ on the 1,770-bet ensemble of §5.7, treating top-quartile log-payoff regret as the harm signal. Brier-regret is mechanically tied to $\varepsilon^{\star 2}$ by the Pythagorean improvement (Corollary B.7); we report it as a secondary check. Log-payoff regret depends on which coordinate resolved YES and is therefore not geometrically tied to $\varepsilon^\star$, making it the genuinely informative discrimination test.

Discrimination. $\varepsilon^\star$ predicts top-quartile log-payoff regret with 5-fold cross-validated AUC 0.775 ± 0.044 (Figure 3a); the geometric Brier-regret check gives CV-AUC 0.974 ± 0.012. Mean realised log-payoff regret stratifies cleanly by $\varepsilon^\star$ quartile (0.000, 0.000, 0.123, 0.221 nats, bottom to top).

Operating thresholds. A high-recall threshold $\tau = 0.15$ catches 91% of harmful bets at 25% alert rate and 3% false-alarm rate; a high-precision threshold $\tau = 0.22$ catches 51% of harmful bets at 14% alert rate and 1.5% FPR (Figure 3b). Both thresholds are stable under 5-fold CV (at $\tau = 0.15$: mean alert rate 25%, mean recall 90%). For deployment we recommend $\tau \approx 0.15$ when repair is cheap and $\tau \approx 0.22$ when escalation is costly. Full operating table in Section O.

5.9. Within-component verification and reproducibility

We verify that the local JCD layer behaves as predicted by Theorem B.1: sorting Paleka cliques by ε, the most-incoherent quartile gains 0.040 Brier at $m=2$ and 0.013 at $m=3$, with no detectable improvement in the lower three quartiles. Per-question Platt calibration shows no such stratification, distinguishing projection from generic calibration. Stacked with Platt, JCD reduces mean Brier by 0.037 on Paleka and 0.046 on Polymarket relative to the raw K-sample mean across every (model, checker) cell ($p < 10^{-4}$ by Diebold–Mariano per model). Code, prompts, sample dumps, ensemble seeds, and the hierarchical-projection pipeline are in Section E; the sequential coherence test in Section D.

6. Discussion and Conclusion

Summary. We characterised *compositional incoherence* – a multi-component LLM agent with locally coherent components emits a globally incoherent joint quote – via the compositional residual $\varepsilon^\star$, a runtime certificate computable from the agent's output and the coupling set alone. A product-structure dichotomy (Theorem 4.2) characterises when local coherence guarantees global coherence; a hierarchical Boyle–Dykstra repair (Theorem B.3) recovers the joint projection deterministically to $\varepsilon^\star \leq 1.5 \times 10^{-16}$. The evaluation of §5 establishes that the failure mode is structural, persists under stronger models and three intuitive LLM-side mitigations, and that among the interventions tested only the geometric repair drives $\varepsilon^\star$ to numerical zero. Compositional incoherence operationalises the "inter-agent misalignment" category of MAST (Cemri et al., 2025) with a scalar certificate ($\varepsilon^\star$) and a deterministic geometric repair, in contrast to the trace-annotation methodology of MAST and the prompting and training-side mitigations of decision-making failures (Schmied et al., 2025).

Implications. Compositional incoherence is detectable from agent outputs alone, without modifying components or accessing internals: any deployed multi-component pipeline emitting probabilistic claims over a known coupling set can compute $\varepsilon^\star$ at runtime and act on it. The empirical pattern – prompt-side mitigations help variably and harm sometimes, geometric projection recovers numerical coherence deterministically – yields a concrete deployment recipe: gate on $\varepsilon^\star > \tau$ and apply the projection only when the certificate fires, avoiding regression on already-coherent quotes. The certificate construction is independent of the chosen norm; KL/I-projection variants apply when the downstream consumer is log-likelihood-based.

Limitations. (i) Theorem 4.2 is existential, with no quantitative magnitude bound on $\varepsilon^\star$; Lemma D.1's envelope is loose at low variance. (ii) Non-linear aggregators and free-form planner-tool traces require inferring the coupling set $\mathcal{C}$, which this paper does not solve. (iii) Retrieval was tested only with a top-5 DuckDuckGo web search per outcome; richer retrieval (multi-source, structured, or query-rewriting) may behave differently. (iv) The LLM-as-aggregator baseline used Claude-Haiku-4.5 at $T=0$ with a single prompt; a more capable aggregator with reasoning enabled and structured-output decoding might close more of the gap. (v) The Polymarket evaluation slice skews toward politics, sports, and awards partitions; broader topical coverage is open. (vi) The τ thresholds of §5.8 are calibrated and cross-validated on the same 1,770-bet pool used to measure regret; a fully independent holdout would be cleaner.

(vii) Operating points assume the coupling set $\mathcal{C}$ is given; inferring $\mathcal{C}$ from free-form traces is left to future work. Code: https://github.com/ananykotawala/composition-incoherence-icml.

References

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

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.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., and Stoica, I. Why do multi-agent LLM systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. arXiv:2503.13657.

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.

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.

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.

Schmied, T., Bornschein, J., Grau-Moya, J., Wulfmeier, M., and Pascanu, R. LLMs are greedy agents: Effects of RL fine-tuning on decision-making abilities. In *EXAIT Workshop at ICML 2025*, 2025. arXiv:2504.16078.

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.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Wornow, M., Garodia, V., Vassalos, V., and Contractor, U. Top of the CLASS: Benchmarking LLM agents on real-world enterprise tasks. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

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.

Zhu, K., Liu, Z., Li, B., Tian, M., Yang, Y., Zhang, J., Han, P., Xie, Q., Cui, F., Zhang, W., Ma, X., Yu, X., Ramesh, G., Wu, J., Liu, Z., Lu, P., Zou, J., and You, J. Where LLM agents fail and how they can learn from failures, 2025. arXiv:2509.25370.

Appendix: Supplementary Material

A. Per-specialist residual attribution

For each ensemble bet with $\varepsilon^\star > 10^{-9}$ ($n_{\text{pos}} = 894$ across the 1,876-bet ensemble of §5.4), we attribute residual mass to each specialist by computing the L_2 norm of the residual vector $(x - \Pi^\star(x))$ restricted to the coordinates that specialist owned. The data are computed from the cached sample dumps of Section E.

Table 2. Per-specialist mean attributed L_2 mass on coordinates each specialist owned, across the 894 positive-residual bets.

Specialist	cliques owned	mean $\|\delta_a\|_2$	median
Claude-Haiku-4.5	506	0.092	0.078
GPT-5.4-mini	605	0.120	0.114
GPT-5.4-nano	491	0.093	0.080
Llama-3.3-70b	464	0.098	0.090

GPT-5.4-mini carries the largest attributed mass per clique, consistent with the frontier-panel finding that the GPT-5.5 upgrade reduces magnitude but not prevalence (§5.6).

B. Definitions, proofs, and remarks moved from §4

Theorem B.1 (Pythagorean Brier improvement (within-component))**.** *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 is the squared distance from coherence; the within-component 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 B.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. The finite-state Fundamental Theorem of Asset Pricing gives no state-independent Dutch book against the ideal projection. With a numerical QP solution, the unit-stake exposure diagnostic 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).*

Theorem B.3 (Hierarchical projection convergence)**.** *Let $\{\Pi_1, \ldots, \Pi_k, \Pi_{\mathcal{C}}\}$ be the family of L_2 projections onto $\{\mathcal{M}_a^\uparrow\}$ and $\mathcal{C}$, with intersection $\mathcal{M}^\star$. The Boyle–Dykstra cyclic iteration (Boyle & Dykstra, 1986) initialised at any $r_0 \in [0,1]^{m^\star}$ converges in L_2 to $\Pi^\star(r_0)$.*

Definition B.4 (Owner-selected aggregation)**.** Fix the joint question set $\mathcal{Q}^\star$. An ownership map $\text{owner} : \mathcal{Q}^\star \to [k]$ 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 $\mathcal{C}$. The aggregator $\mathcal{A}$ selects the coordinate owned by $\text{owner}(j)$. Aggregation only *selects*; coupling *enforces agreement*.

Remark B.5 (Scope and the averaging aggregator). Theorem 4.2 characterises owner-selected coordinate aggregation. For an averaging aggregator $\mathcal{A}_{\text{avg}}$ that emits the mean of per-component quotes on shared logical coordinates, the disjoint-coordinate case (which covers all experiments in §5) coincides with owner-selection and inherits Theorem 4.2 verbatim. On a *shared* coordinate j quoted as α by component a and β by component b, $\mathcal{A}_{\text{avg}}$ emits $(\alpha+\beta)/2$ – exactly the L_2 projection of (α, β) onto the equality coupling $\{p_{a,j} = p_{b,j}\}$; averaging *is* the local repair on shared-coordinate equality. Hence Theorem B.3's Boyle–Dykstra cycle specialises to averaging when the only coupling is shared-coord equality, and to general $\Pi^\star$ otherwise. Log-pooling and other non-coordinate aggregators are subject to distinct dichotomies under KL geometry.

Proof of Theorem 4.2 (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 minimisation separates into k independent local minimisations. 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; thus $\varepsilon^\star \equiv 0$ (Bauschke & Combettes, 2017, Prop. 29.3). □

Proof of Theorem 4.2 (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 $\hat{p}_i^{(a)} := r_{j(a,i)}$ where $j(a,i)$ is the joint index that component a's i-th local coordinate maps to under Definition B.4; 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, $\varepsilon^\star = \|r - \Pi^\star(r)\|_2 > 0$. □

Corollary B.6 (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. Under the unit-stake LMSR exposure statistic (Hanson, 2003),* $\mathrm{Exposure}^\star \le \sqrt{m^\star}\,\varepsilon^\star$. *Per-component exposure is zero by construction up to QP tolerance.*

Proof. If $x \notin \mathcal{M}^\star$, de Finetti's finite-state FTAP yields a portfolio with nonnegative payoff in every coherent state and positive price discrepancy. For our relation classes the unit-stake exposure is the positive halfspace violation $h(x) = (\langle a, x\rangle - b)_+ \le \|a\|_2 \|x - \Pi^\star(x)\|_2$; $\|a\|_2 \le \sqrt{m^\star}$ in $m^\star$ joint coordinates yields the bound. Per-component residuals to $\mathcal{M}_a$ are zero by construction. □

Corollary B.7 (System-level Pythagorean Brier improvement)**.** *For any composed quote r and joint truth $p^\star \in \mathcal{M}^\star$,* $\|\Pi^\star(r) - p^\star\|_2^2 \le \|r - p^\star\|_2^2 - (\varepsilon^\star)^2$.

Proof. Hilbert projection theorem applied to $\Pi^\star$ on $\mathcal{M}^\star$. □

Proof of Proposition 4.3. By the variational characterisation of L_2 projection onto a closed convex set, $\varepsilon^\star = \|x - \Pi^\star(x)\|_2 \le \|x - r\|_2$ for any $r \in \mathcal{M}^\star$, where $x = \mathcal{A}(\Pi_1 \hat{p}^{(1)}, \ldots, \Pi_k \hat{p}^{(k)})$. □

Sketch of Theorem B.3. $\mathcal{M}^\star = \left(\bigcap_a \mathcal{M}_a^\uparrow\right) \cap \mathcal{C}$; each $\Pi_\bullet$ projects onto a closed convex set. 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. □

C. Closed-form local projections

Negation (m=2, $r_1 + r_2 = 1$)**.** $\Pi_{\mathcal{M}_C}(\hat{p}) = \left(\frac{1}{2}(1 + \hat{p}_1 - \hat{p}_2),\ \frac{1}{2}(1 - \hat{p}_1 + \hat{p}_2)\right)$; 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 (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

The compositional theory of Section 4 controls coherence at a single time step. For long-horizon deployments, the residual stream $(\varepsilon_t)_{t \ge 1}$ admits a complementary anytime-valid test. Under coordinate-selection aggregation with disjoint owners, changing one Bernoulli sample changes $\varepsilon_t^{\star 2}$ by at most $2/\min_a K_{a,t}$ on the affected coordinate; Lemma D.1 therefore inherits at the system level with m_t replaced by $m_t^\star$ and K_t by $\min_a K_{a,t}$.

Setup. A clique stream $(C_t)_{t\geq 1}$ 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 $\mathbb{E}[\varepsilon_t^2 \mid \mathcal{F}_{t-1}] \leq m_t/(4K_t)$, and for every $\lambda > 0$, $\mathbb{E}\big[\exp\big(\lambda(\varepsilon_t^2 - m_t/(4K_t))\big) \mid \mathcal{F}_{t-1}\big] \leq \exp\big(\lambda^2 m_t/(2K_t)\big)$.*

Proof. $\varepsilon_t \leq \|\hat{p}_t - p_F^{(t)}\|_2$ 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 $\mathbb{E}[\exp(\lambda(Z_t - \mathbb{E}Z_t))] \leq \exp(\lambda^2 m_t/(2K_t))$, and $\mathbb{E}Z_t \leq m_t/(4K_t)$ 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, $\mathbb{P}\big(\sup_{t\leq\tau} E_t(\lambda) \geq 1/\alpha \mid H_0\big) \leq \alpha$. The decision rule $\tau_\alpha := \inf\{t : E_t \geq 1/\alpha\}$ 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, $\mathbb{E}[E_t/E_{t-1} \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2 m_t/(2K_t))\exp(-\lambda^2 m_t/(2K_t)) = 1$, 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 4), retaining anytime-valid coverage by linearity (Ramdas et al., 2023).

Empirical demonstration. Figure 4 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 4 after replacing the per-component envelope by the corresponding envelope for the agent's aggregation rule.

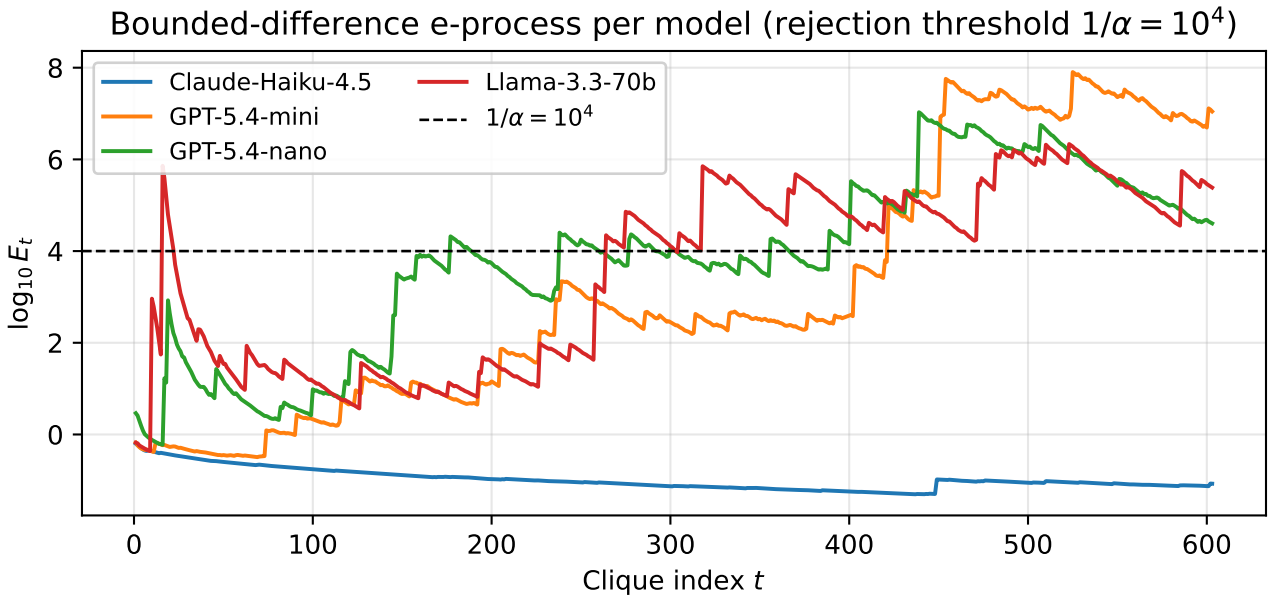

Figure 4. **Bounded-difference e-process E_t on per-component residual streams (Theorem D.2).** Crossing $1/\alpha = 10^4$ at t=16 (Llama), t=177 (GPT-nano), t=422 (GPT-mini); Claude-Haiku stays below 1.

E. Clique-mining pipeline and reproducibility

Paleka. Tuples are pre-mined by the original authors: each JSONL record is a single $(\text{checker}, \{Q_i\}, \text{relation})$ 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 B.3 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 the resolution-status verification of the source benchmark (Paleka et al., 2025).

Reproducibility manifest. Table 3 records the exact configuration used to produce every number in the paper. The codebase (`https://github.com/ananykotawala/composition-incoherence-icml`) contains the experiment scripts (`e1`–`e6`, `t11`, `t12` corresponding to the experiments described in §5 and Appendix), the cached JSON outputs used to produce every figure and table, the sample dumps, per-clique JCD residuals, ensemble-assignment seeds, and the hierarchical-projection pipeline. Each script is self-contained and parameter-free; the downstream-regret, trivial-vs-Boyle–Dykstra, and disagreement-correlation analyses reproduce without LLM API access (operating only on the cached sample dumps).

Table 3. Reproducibility manifest. N values include ensemble-assignment seeds where applicable.

Item	Value
Specialist model IDs	`claude-haiku-4-5-20251001`, `gpt-5.4-mini`, `gpt-5.4-nano`, `llama-3.3-70b-versatile`
Snapshot date	30 April 2026
K samples / question	8
Temperature	0.7
Sampling prompts	JSONL artifact in code release
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. Structured-routing replication

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

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.

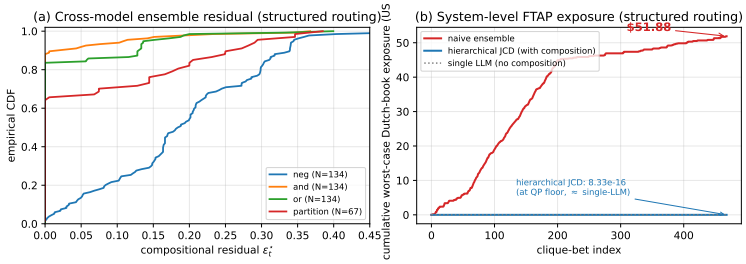

Figure 5. **Structured-routing replication of Figure 2.** The same certificate is positive with a different relation breakdown but comparable per-bet exposure ($0.11/bet vs. $0.13/bet under random assignment). The lower cumulative endpoint reflects fewer bets under structured routing, N=469 vs. N=1,876, not lower per-bet exposure.

G. Worst-case failure modes

Table 4 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 4. 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

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

We report all 100 live-style partition forecasting questions used in the planner-style routing simulation (Section 5). 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 5. 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

(continued on next page)

(continued from previous page)

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

(continued on next page)

(continued from previous page)

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

I. Trivial vs. hierarchical projection: per-relation gaps

We verify that on every one of the $1{,}876$ random-assignment ensemble bets, the per-relation closed-form projection of Section C coincides with the hierarchical Boyle–Dykstra output to within the gaps below (computed on the composed quote under 4 random-assignment seeds; master seed 0).

Table 6. Per-relation maximum $\|\Pi^{\text{closed}} - \Pi^{\text{JCD}}\|_\infty$ over the $1{,}876$ ensemble bets. Negation and partition match to numerical precision; conjunction/disjunction require the cyclic projection to recover the exact L_2 projection because the Fréchet polytope is not a Cartesian product of half-spaces.

Relation	n bets	max gap	mean gap
`neg`	536	1.1×10^{-16}	1.7×10^{-17}
`partition`	268	0 (exact)	0
`and`	536	7.8×10^{-2}	1.3×10^{-3}
`or`	536	8.8×10^{-2}	3.2×10^{-3}

J. Inter-LLM disagreement and the residual

For each ensemble bet (master seed 0, 4 random-assignment seeds), we compute (i) the disagreement bound of Proposition 4.3, taken as $\min_\beta \|x - \Pi_\beta(\hat{p}^{(\beta)})\|_2$ over reference models β, and (ii) a simple pairwise-max disagreement $\max_{a,b} \|\Pi_a(\hat{p}^{(a)}) - \Pi_b(\hat{p}^{(b)})\|_2$. Restricted to positive-residual cliques, the bound is highly predictive of $\varepsilon^\star$ across all relations, and even the plain pairwise-max disagreement explains a moderate share of variance via linear regression.

Table 7. Spearman rank correlation between disagreement-based predictors and $\varepsilon^\star$, restricted to cliques with $\varepsilon^\star{>}0$.

Relation	$n_{\varepsilon>0}$	$\rho(\text{bound}, \varepsilon^\star)$	$\rho(\text{pair}, \varepsilon^\star)$	$R^2(\text{pair}, \varepsilon^\star)$
`neg`	343	0.97	0.51	0.29
`partition`	148	0.83	0.31	0.09
`or`	215	0.79	0.49	0.26
`and`	188	0.70	0.28	0.08

K. Downstream-decision regret on resolved cliques

For each of the $1{,}770$ ensemble bets retained after dropping partition cliques with per-model size misalignment, we converted the composed quote $\hat{p}^\star$ into a unit-budget proportional allocation $w_i = \max(\hat{p}^\star_i, 0)/\sum_j \max(\hat{p}^\star_j, 0)$ and computed Brier against the binary resolution vector and log-payoff $\log w_{\text{winner}}$ on the $1{,}250$ bets where exactly one coordinate resolved YES. Confidence intervals are 95% paired bootstrap.

Table 8. Downstream regret of three regimes on the $1{,}770$-bet ensemble. "oracle" is a single-LLM JCD forecast on the full clique. Brier is the sum of squared coord-wise errors against the binary resolution vector; lower is better. Log-payoff is in nats; higher is better.

Regime	mean Brier	mean log-payoff	n
naive composition	0.681	-1.044 (when winner)	1,770
hierarchical JCD	0.663	-0.929	1,770
single-LLM oracle	0.664	-0.978	1,770
Δ(JCD$-$naive)	-0.018 $[-0.021, -0.015]$	$+0.115$ $[+0.056, +0.187]$	paired
Δ(oracle$-$naive)	-0.017 $[-0.035, -0.000]$	$+0.043$ $[-0.045, +0.125]$	paired

L. Partition K-sweep across LLMs: structural vs sampling

To rule out that the headline "94% of partitions positive" result of §5 is finite-K sampling noise, we re-elicit K=16 verbalised samples per (model, outcome) on a seeded subset of 12 planner-routing partitions (Section H) for the three specialists with available quota at run time (Claude-Haiku-4.5, GPT-5.4-mini, GPT-5.4-nano; Llama-3.3-70b's host hit its daily-token budget at elicitation time). For each $K_{\text{target}} \in \{4, 8, 16\}$ and each of 4 random-assignment seeds, we sub-sample the first K_{target} samples per (model, outcome), assemble a composed quote under random outcome-to-specialist assignment, and compute $\varepsilon^\star$.

Table 9. K-sweep on 12 partitions (n=48 bets per K). Mean $\varepsilon^\star$ is essentially flat; the $\varepsilon^\star$>0.05 rate is $\geq 85\%$ at every K.

K	n	mean $\varepsilon^\star$	median $\varepsilon^\star$	$\Pr(\varepsilon^\star{>}0.05)$
4	48	0.201	0.174	0.854
8	48	0.203	0.178	0.896
16	48	0.210	0.174	0.896

If the failure mode were finite-K noise, mean $\varepsilon^\star$ would shrink at the $1/\sqrt{K}$ rate; instead it is flat (slight increase from K=4 to K=16 is within seed-induced variation). The dominant residual mass is structural cross-component disagreement, not finite-sample fluctuation.

Robustness check including Llama. A subsequent re-run added Llama-3.3-70b back to the panel (when host RPM allowance permitted). Per-minute throttling prevented obtaining K=16 for many bets, so K=16 is dropped; the partial 4-model results are mean $\varepsilon^\star = 0.274$ at K=4 (n=28 bets, after dropping bets where Llama failed to produce K samples) and 0.288 at K=8 (n=20). Both are *higher* than the 3-model values above (0.20–0.21), consistent with Llama contributing additional inter-component disagreement; positive-residual rate is 96% (K=4) and 100% (K=8). The flat-vs-K pattern is preserved.

M. Causal-mechanism (coupling-visibility) experiment

This appendix supports §5.3.

Sample-of-partitions. The 20 partitions are the *full* planner-harness slate of §5.2; the spec "20 highest-$\varepsilon^\star$" selects every partition because the harness has $\varepsilon^\star > 0$ on 20/20.

Routing. Held fixed across BLIND and INFORMED at the planner-harness's chosen specialist-to-outcome assignment per partition. (The original planner harness has no random-assignment seeds; the four "seeds" here are independent K=8 sampling rounds per (partition, condition), giving four paired $\varepsilon^\star$ values per partition.)

Prompts. Verbatim from the experiment specification: BLIND is a single-Bernoulli question; INFORMED includes the partition label, all sibling outcomes, the $\sum_i p_i = 1$ constraint, and the peers' BLIND quotes on the other outcomes. INFORMED uses BLIND quotes from the same seed as context, so the two conditions are tightly paired within a (partition, seed) cell.

Generation budget. Claude max_tokens raised $64 \to 512$, and Llama max_tokens raised $16 \to 512$, because the longer INFORMED prompt induces a chain-of-thought preamble before the number; the smaller budgets clip before the emission. This is a generation-side parameter only; prompts are verbatim from the spec.

Llama specialist substitution. The original BLIND run hit a per-day token-budget limit on `llama-3.3-70b-versatile` mid-experiment, producing rate-limit failures on the 9 partitions whose planner-chosen

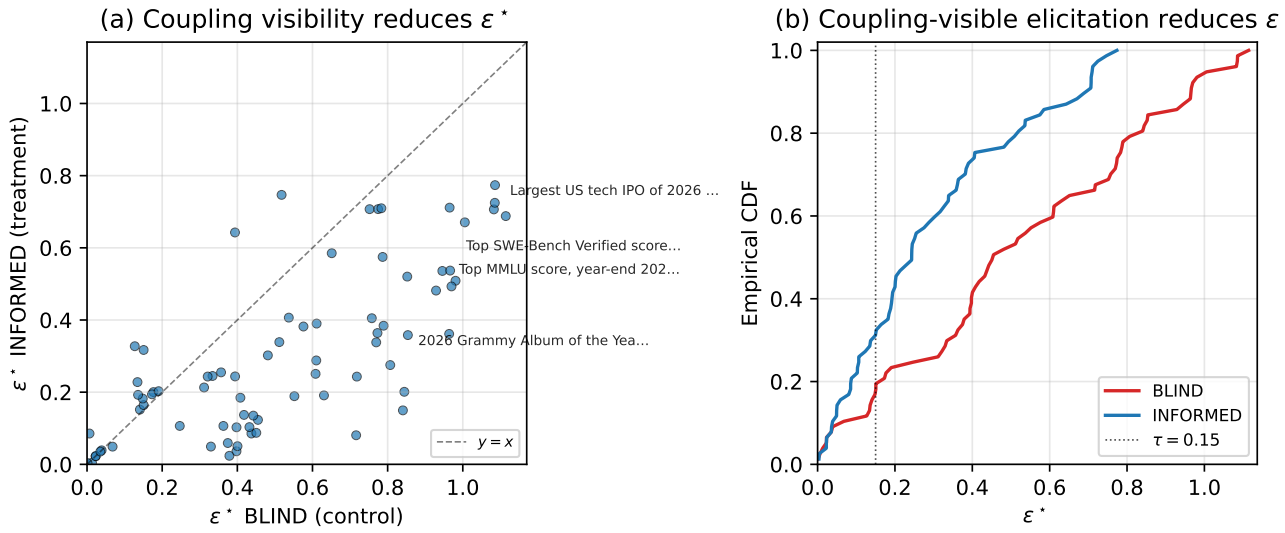

Figure 6. **Coupling-visibility causal intervention (§5.3).** Each point is one (partition, seed) pair. *(a)* Paired scatter, BLIND on x-axis, INFORMED on y-axis; points below the $y = x$ diagonal indicate the intervention reduced $\varepsilon^\star$ ($61/77$ paired observations). The four largest-BLIND partitions are annotated by name. *(b)* Empirical CDFs under each condition. Vertical line at $\tau = 0.15$, the high-recall gating threshold of §5.8.

routing assigns at least one outcome to Llama. We substituted `Llama-4-Maverick-17B-128E-Instruct-FP8` for those 9 partitions (re-elicited *all* four seeds $\times$ both conditions for the Llama-owned outcomes, 576 additional calls). Each partition is internally consistent: a single Llama variant across all of its (seed, condition) cells. The remaining 11 partitions retain Llama-3.3-70b throughout. Logged in the per-call manifest of the supplementary release; the model field reads either `llama-3.3-70b` or `llama-4-maverick-17b-128e-fp8`.

Statistical test. Wilcoxon signed-rank, two-sided, paired by (partition, seed). Bootstrap CI is 5,000-iteration paired bootstrap over the same pairs. Final $n_{\text{pairs}} = 77$ from $20 \times 4 = 80$ targeted cells; 3 cells lost to terminal parse failures after both the generation-budget bump and the Llama specialist substitution.

N. Frontier-model panel detail

Specialists and routing. Each model is queried through a single hosted inference platform for that model, all behind one API client for uniform tooling. GPT-5.5 is queried with `reasoning_effort='none'`, `max_completion_tokens=64`, `temperature=0.7`; the other three with `max_tokens=32`, `temperature=0.7`. We disable GPT-5.5's reasoning trace to avoid spending the entire token budget on chain-of-thought, ensuring apples-to-apples comparison with the three non-reasoning models.

Cliques. The 67 partition cliques are sampled (seed 0) from the Polymarket multi-candidate events in the gamma-api cache that have exactly one resolved YES outcome and per-market question text (i.e. not boilerplate resolution criteria). All 67 resolved strictly after the model snapshot date (April 30, 2026) and before our analysis date, matching the leakage-control protocol of §5.1.

Sample sizes. Full per-LLM coverage on $67/67$ partition cliques; 9,632 LLM calls; 268 ensemble bets (67 cliques $\times$ 4 random-assignment seeds), all with a single resolved winner.

Table 11. Frontier vs. mid-tier panel (n=268 partition bets each). 95% paired-bootstrap CIs over (clique, seed) pairs; the mid-tier numbers are the partition-only projection of §5.4/§5.7.

Table 10. Coupling-visibility intervention: BLIND specialists (no partition context) vs INFORMED specialists (full partition context + peer quotes), on the 20 highest-$\varepsilon^\star$ partitions of the planner harness (§5.2). $\varepsilon^\star$ values are mean ± std over 4 sampling seeds; $\Delta = \varepsilon^\star_{\text{BLIND}} - \varepsilon^\star_{\text{INFORMED}}$. Sign: ↓ improved, ∘ unchanged, ↑ worsened.

Partition	m	$\varepsilon^\star_{\text{BLIND}}$	$\varepsilon^\star_{\text{INFORMED}}$	Δ	sign
2026 Grammy Album of the Year genre	4	0.848±0.005	0.307±0.145	+0.541	↓
2026 Booker Prize winner nationality	3	0.668±0.050	0.191±0.068	+0.477	↓
Top MMLU score, year-end 2026	4	0.956±0.020	0.505±0.020	+0.451	↓
Top SWE-Bench Verified score, year-end 202	4	0.975±0.018	0.570±0.137	+0.405	↓
Largest US tech IPO of 2026 by company typ	4	1.092±0.013	0.723±0.032	+0.369	↓
2026 Academy Award Best Picture winner typ	3	0.784±0.017	0.440±0.191	+0.344	↓
2026 FIFA World Cup winner confederation	3	0.390±0.011	0.066±0.028	+0.324	↓
2026 US Senate majority	3	0.439±0.018	0.133±0.035	+0.306	↓
2026 US House majority	3	0.420±0.017	0.142±0.061	+0.278	↓
2026 Wimbledon men's singles winner	4	0.783±0.006	0.508±0.143	+0.275	↓
Next UK general election year	4	0.394±0.056	0.127±0.104	+0.267	↓
Top gold-medal nation, 2028 Summer Olympic	3	0.580±0.024	0.320±0.093	+0.259	↓
2026 Pulitzer Prize Fiction winner setting	3	0.693±0.064	0.496±0.161	+0.197	↓
Next FOMC decision (June 2026 meeting)	3	0.331±0.017	0.239±0.016	+0.092	↓
2026 H1 BTC return bucket	3	0.236±0.093	0.165±0.113	+0.071	↓
2026 NBA Finals winner conference	2	0.030±0.007	0.030±0.007	+0.000	∘
2026 US CPI year-over-year, December print	4	0.022±0.027	0.036±0.034	−0.013	↑
US unemployment rate, June 2026	4	0.160±0.015	0.178±0.020	−0.017	↑
2026 H1 SP500 return bucket	4	0.490±0.056	0.534±0.167	−0.044	↑
2028 US Presidential winner	3	0.142±0.007	0.230±0.053	−0.088	↑
Mean across 20	—	0.519	0.298	**+0.221** [+0.173, +0.270]	—

Metric	Frontier	Mid-tier
$\Pr(\varepsilon^\star > 0)$	0.978	0.940
$\Pr(\varepsilon^\star > 0.05)$	0.433	—
$\Pr(\varepsilon^\star > 0.10)$	0.257	—
mean $\varepsilon^\star$	0.072	0.118
median $\varepsilon^\star$	0.042	—
ΔBrier (JCD)	−0.012 [−0.015, −0.009]	−0.018 [−0.021, −0.015]
Δlog-p (JCD)	+0.107 [−0.12, +0.36]	+0.60 [+0.15, +1.15]
Δlog-p (oracle)	−0.051 [−0.42, +0.32]	≈ 0

O. Threshold-calibration / ROC details

We construct the harm-prediction problem as follows. For each of the 1,770 bets in the random-assignment ensemble of §5.7, we record $\varepsilon^\star$ on the composed quote, Brier(naive) − Brier(JCD) (Brier regret), and on the 1,247 bets with a single resolved winner, log-payoff(JCD) − log-payoff(naive) (log-payoff regret). The positive class is top-quartile of either regret signal; the predictor is $\varepsilon^\star$.

Table 12. Mean realised regret by $\varepsilon^\star$ quartile ($n_{Q1=0}$=0 for the bottom-quartile log-payoff column because all such bets have JCD ≡ naive).

$\varepsilon^\star$ quartile	range	$\Delta\bar{\text{Brier}}$	$\Delta\bar{\text{log-p}}$	n
Q1	[0.000, 0.000]	0.000	0.000	∼440
Q2	(0.000, 0.004]	0.000	0.000	∼440
Q3	(0.004, 0.154]	0.012	0.123	∼440
Q4	(0.154, 0.545]	0.054	0.221	∼440

Table 13. Operating points for the $\varepsilon^\star$ gate, computed in-sample then verified by 5-fold CV (mean shown in parentheses where it differs from the in-sample number).

Mode	τ	alert rate	harm capture	FPR
high-recall ($\geq 90\%$ capture)	0.153	25.1%	91.0% (90.4%)	3.1%
balanced	0.180	19.5%	73.5%	2.3%
high-precision ($\geq 50\%$ capture)	0.222	13.8%	50.7%	1.5%
conservative	0.300	5.6%	22.7%	0.7%

Cross-validated AUC (mean $\pm$ std across 5 folds, harm threshold and ROC computed on disjoint train/test halves): 0.974 ± 0.012 for top-quartile Brier-regret prediction (Pythagorean lower bound geometrically present); 0.775 ± 0.044 for top-quartile log-payoff-regret prediction (no geometric tie). Spearman $\rho(\varepsilon^\star, \text{Brier regret}) = 0.829$; $\rho(\varepsilon^\star, |\text{log-p regret}|)$ on positive-residual bets is significantly positive (the near-zero per-bet correlation with signed log-p regret reflects that JCD's mass redistribution can either help or hurt a given bet depending on the winning coordinate, even as it reduces *magnitude* of regret in expectation).

P. Allocation-rule robustness for the regret experiment

The proportional rule of §5.7 is one of several reasonable ways for a downstream bettor to convert an agent's quote into a unit-budget allocation. We check the regret result under three rules:

- **Proportional.** $w_i = \max(p_i, 0)/\sum_j \max(p_j, 0)$.
- **Truncated Kelly.** $w_i = \max(p_i, 0)$ when $\sum_j p_j \leq 1$ with $(1 - \sum p)/m$ added uniformly across outcomes; proportional truncation when $\sum p > 1$.
- **Max-entropy.** $\arg\max_w H(w)$ subject to $\sum w_i = 1$, $w_i \geq \max(p_i, 0)$ when feasible (sum of floors ≤ 1); fall back to proportional otherwise.

Table 14. JCD vs. naive log-payoff under three allocation rules (paired bootstrap, $n = 1{,}252$ bets). The proportional rule preserves the agent's incoherence as allocation slack; the stricter rules *themselves* coherentise (top up to sum 1 when $\sum p < 1$, scale down when $\sum p > 1$), absorbing the agent's incoherence into the bettor's behaviour. Brier is allocation-independent: -0.018 $[-0.021, -0.015]$ across all rules.

Rule	$\Delta_{\text{naive}} \to$jcd	95% CI
proportional	$+0.168$	$[+0.083, +0.263]$
truncated Kelly	$+0.006$	$[-0.003, +0.017]$
max-entropy	$+0.006$	$[-0.003, +0.017]$

The Brier result is the proper-scoring-rule conclusion (it evaluates the agent's emitted quote directly without an intermediate bettor). The proportional rule's $+0.168$ nats per bet is the clearest behavioural-harm measure: when the bettor treats the agent's quote as-is and merely normalises, the agent's incoherence costs the bettor measurable log-wealth. Stricter allocation rules effectively perform their own coherentisation, absorbing the agent's incoherence into the bettor's strategy – which works in the partition case, but generalises poorly to non-partition coupling (negation, conjunction/disjunction, implication ladders) where "top-up to sum 1" is not the right operation.

Q. LLM-as-aggregator baseline (§5.5): protocol and failure modes

For each of the 100 partitions in Section H, we hand the assembled marginals (with the assigned-specialist label per outcome), the partition label, and the resolution date to Claude-Haiku-4.5 at T=0 and ask for a coherent partition vector. The exact prompt is:

> "You are an aggregator agent. You receive per-outcome marginal probability estimates for a multi-candidate forecasting partition *[...]* Your job: reconcile these marginals into a SINGLE coherent partition probability vector. Constraints on your output: *[...]* probabilities must sum to 1. *[...]* Output ONLY a comma-separated list of $\{n_{\text{outcomes}}\}$ probabilities in the SAME ORDER as the outcomes above."

The LLM aggregator achieves $\varepsilon^\star < 10^{-3}$ on $81/100$ partitions and $\varepsilon^\star < 10^{-2}$ on $83/100$. On the remaining 17–19, the aggregator emits a non-coherent vector despite the explicit instruction; the seven *regressions* ($\varepsilon^\star_{\text{LLM}} > \varepsilon^\star_{\text{naive}}$) include "2026 Nobel Physics laureate field" (input sum 1.32 $\to$ output sum 2.00), "Largest US tech IPO of 2026" (input sum 1.66 $\to$

output sum 1.65, with $\varepsilon^\star = 0.325$), and several partitions where the aggregator returned the input verbatim. JCD achieves $\varepsilon^\star \leq 1.5 \times 10^{-16}$ on all 100 partitions deterministically and without an LLM call. Total cost of the LLM-aggregator experiment: 100 calls, ∼90 s wall-clock, <$0.50.

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

We re-run the first 30 partitions of the case study under a tool-augmented protocol (§5.5). 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 $\Delta\varepsilon^\star = \varepsilon_t^\star - \varepsilon_b^\star$ is a direct A/B. Total cost: ≈720 LLM calls + 90 search calls, ≈$2, ≈25 minutes wall-clock.

Aggregate: mean $\varepsilon_b^\star = 0.260 \rightarrow \varepsilon_t^\star = 0.283$; 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(their bucket)≈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.

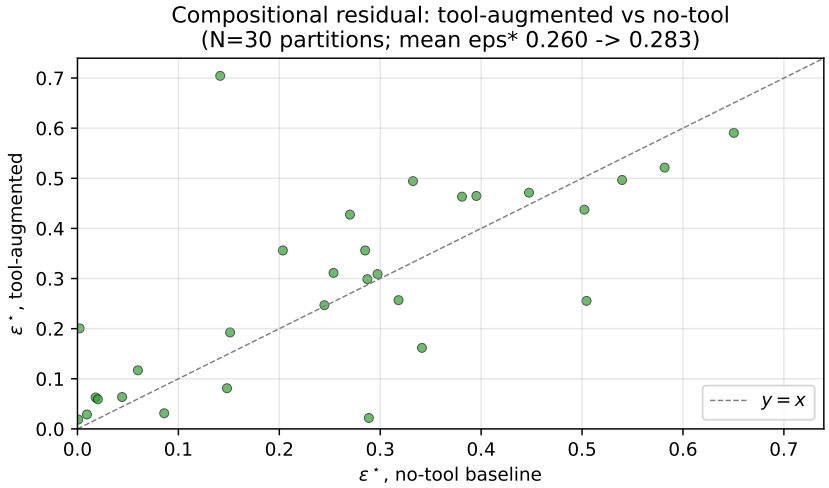

Figure 7. **Tool augmentation does not eliminate the compositional residual.** Each point is one of the 30 matched partitions; x-axis is $\varepsilon^\star$ in the no-tool baseline, y-axis is $\varepsilon^\star$ when each specialist receives top-five DuckDuckGo snippets for its assigned outcome before 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 ↑

(continued on next page)

(continued)

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