

What Ensemble Brier Scores Hide: Evidence Routing at a Fixed Article-Exposure Budget in LLM Prediction-Market Forecasting

Avinash Goutham Aluguvelly

University of Texas at Austin

algo@utexas.edu

Abstract

We ask whether giving different evidence to different LLM forecasters creates useful diversity when the total article-exposure count is held fixed. Across 708 resolved prediction-market contracts, we compare six-member ensembles whose forecasters either share the same two-article pair or are distributed evenly across three pairs, with both conditions using twelve article exposures. On 609 pre-market-end contracts, evidence routing reliably increases within-ensemble dispersion for GPT-4.1-mini, Qwen3-30B, and OLMo-3.1-32B. An exact Brier-score decomposition separates changes in average member Brier, dispersion, and ensemble Brier, while an evidence-pair attribution analysis attributes nearly all of the dispersion increase to differences between assigned article pairs rather than repeated sampling. Greater dispersion, however, does not automatically improve the ensemble: only Qwen’s reduction in ensemble Brier survives multiplicity correction, while GPT and OLMo show no statistically robust net improvement. Post-hoc controls show that more balanced routing further increases dispersion at a fixed evidence union, while an alternative Qwen readout changes channel attribution without establishing a different net effect. Evidence routing therefore consistently increases forecast diversity in this setting, but its value depends on whether that diversity outweighs any loss in individual forecast quality.

1 Introduction

Prediction markets aggregate beliefs about uncertain events into probabilistic market prices. An LLM forecasting system built to support them must decide not only which documents to retrieve, but how to allocate those documents across repeated model calls. A natural strategy gives different evidence to different ensemble members and averages their probabilities (Li et al., 2026). Yet disagreement is not automatically useful: broader evidence

coverage may diversify forecast errors while simultaneously weakening individual forecasts.

This trade-off is hidden by the aggregate score. Under squared error, an ensemble’s Brier score equals its average member Brier minus its within-ensemble dispersion. Routing can make the forecasters disagree more while also making their individual predictions worse. Because these effects pull the ensemble score in opposite directions, they can cancel. We ask whether routing different evidence to parallel LLM forecasters creates useful diversity when the model, prompt, ensemble size, per-member article count, and total article-exposure count are held fixed.

We study 708 resolved prediction-market contracts across three served model endpoints. In the primary comparison, six independent neutral forecasters either receive the same two-article pair or are divided evenly across the three possible pairs. Both conditions give every member two articles and use twelve article exposures across the ensemble. With three articles A_1 , A_2 , and A_3 , for example, the shared ensemble gives every member the same pair, whereas the routed ensemble assigns two members to each of (A_1, A_2) , (A_1, A_3) , and (A_2, A_3) . Routing consequently expands panel-level coverage from two articles to three; it does not hold the union of available information fixed.

On the 609 contracts whose snapshots precede market end, routing increases within-ensemble dispersion for all three endpoints, with nearly all of the increase attributable to differences between assigned article pairs rather than repeated sampling. Greater dispersion does not uniformly translate into better ensemble forecasts: only Qwen’s ensemble-Brier improvement survives multiplicity correction. Post-hoc analyses show that more balanced routing further increases dispersion even when panel-level evidence coverage is held fixed, and probe whether the channel attribution depends on the probability readout.

Our contributions are:

1. a controlled routed-versus-shared comparison at a fixed article-exposure count;
2. a paired decomposition of routing effects into average member Brier, dispersion, and ensemble Brier across three served endpoints; and
3. evidence that routing creates structured forecast diversity, together with the qualification that greater diversity need not improve the ensemble.

2 Related work

LLMs can produce competitive forecasts when retrieval, reasoning, aggregation, and calibration are combined (Halawi et al., 2024). Forecast-Bench standardizes dynamic forecasting evaluation (Karger et al., 2024), and PolyBench extends evaluation to live prediction-market forecasting and trading (Cheng et al., 2026). Other work treats multiple LLMs as a “silicon crowd” (Schoeneger et al., 2024) or directly optimizes the trade-off between accuracy and diversity in LLM mixtures (Aitchison et al., 2026). Our unit of analysis is instead a single fixed served endpoint, within which we vary the routing of retrieved evidence.

Information asymmetry has also been used to improve multi-agent deliberation. Li et al. (2026) use an ensemble-error analysis to motivate public/private evidence routing and measure inter-agent variance before and after communication; their InfoDelphi system combines BM25-based routing, iterative rationale exchange, and confidence-weighted logit aggregation. We do not propose a new decomposition. The exact squared-error identity is the ambiguity decomposition of Krogh and Vedelsby (1994), and we use it as the estimand in a matched, noncommunicating, arithmetic-mean experiment that separates member, dispersion, and net channels across served endpoints and readouts.

Verbalized probabilities can also be low-cardinality and rounding-biased (Beyazit et al., 2025), and model confidence depends on the readout protocol used to elicit it (Kim and Kang, 2026). We test whether our routing decomposition survives a balanced label-logit readout, instead of treating a verbalized number as a protocol-free model belief.

3 Analysis framing

For market i , let p_{ij} be member j ’s probability, $y_i \in \{0, 1\}$ the outcome, and $\bar{p}_i$ the arithmetic mean, which we use as the ensemble forecast. The Brier score is mean squared error, and lower is better. For N markets and k members,

$$\underbrace{\frac{1}{N} \sum_i (\bar{p}_i - y_i)^2}_{\text{ensemble Brier}} = \underbrace{\frac{1}{Nk} \sum_{i,j} (p_{ij} - y_i)^2}_{\text{average member Brier}} - \underbrace{\frac{1}{Nk} \sum_{i,j} (p_{ij} - \bar{p}_i)^2}_{\text{within-ensemble dispersion}}. \quad (1)$$

This identity is exact (Krogh and Vedelsby, 1994). The three terms have distinct interpretations. Average member Brier is the average squared error of the individual forecasts, so its change tells us whether routing improves or weakens individual forecast accuracy. Dispersion is their variance around the ensemble forecast and quantifies how much the ensemble gains from averaging forecasts that disagree. Ensemble Brier is the squared error of the averaged forecast, so its change tells us whether the final combined forecast improves.

For an intervention, $\Delta E = \Delta M - \Delta D$. Routing therefore improves the ensemble only when the gain from greater dispersion outweighs any deterioration in individual forecast accuracy. Dispersion—also called the ambiguity/diversity term—acts like a diversification benefit and is measured in Brier units, but it is not by itself evidence that forecasters possess independent or useful information. We call a positive ΔM a member-Brier cost without assuming its sign in advance.

4 Experimental setup

Data and temporal filter. PolyBench provides point-in-time snapshots of Polymarket binary markets, resolution criteria, and three retrieved news articles (Cheng et al., 2026). These snapshots were collected from February 6 to 12, 2026. The dated GPT and Qwen identifiers predate that window, and OLMo’s model card reports a December 2024 data cutoff (Team OLMo, 2025), which reduces our concern about parametric outcome recall. The source contains 11 879 resolved markets.

We apply deterministic filters: binary YES/NO resolution, non-empty criteria, exactly three articles, no forced even settlement, snapshot before official resolution, and one market per parent event. The resulting 708 markets have a 44.6% YES base

ID	Member configuration	Article allocation	Analysis role
C0	1 neutral forecast	$(A_1, A_2, A_3) \times 1$	Secondary
C1	6 neutral members	selected pair $\times 6$	Primary shared
C3	6 neutral members	$(A_1, A_2) \times 2; (A_1, A_3) \times 2; (A_2, A_3) \times 2$	Primary routed
C5	6 neutral members	$(A_1, A_2, A_3) \times 6$	Secondary

Table 1: Primary conditions and practical comparators. A_1 , A_2 , and A_3 denote the three retrieved articles. The shared pair in C1 is selected deterministically for each market. C3 versus C1 is primary; it holds the endpoint, prompt, ensemble size, per-member article count, decoding request, and arithmetic aggregation fixed while reallocating evidence exposures. Exploratory role-conditioned conditions are documented in Appendix B.

rate and climatology Brier 0.2471. When multiple eligible markets share a parent event, we retain the highest-volume market, breaking exact volume ties by ascending stored market id. The append-only manifest preserves that deterministic admission order. For a post-hoc domain sensitivity, we use the overlapping parent-event tags stored in the February 6 PolyBench mirror; these metadata were frozen before the first study snapshot, and 86 strict markets carry the exact tag *Finance*.

Official resolution can post after an outcome is already public. The median snapshot-to-resolution gap is 3.2 days, but 93 snapshots occur at or after the market’s stated end, and 6 markets lack a usable end time. We therefore designate the 609 snapshots that precede market end as our primary sample and report all 708 as a sensitivity check. PolyBench does not provide reliable per-article timestamps, so even this filter cannot establish that every retrieved article predates the outcome. We also use a contemporaneous market probability as a descriptive external baseline. We parse PolyBench’s `market_snapshots.market_prices` vector at the exact retained market id and timestamp, select the entry aligned with the case-insensitive YES label in `markets.outcomes`, and apply neither clipping nor normalization. One empty vector leaves 608 parseable strict markets.

Conditions and evidence budget. The primary C1 and C3 ensembles each have 6 members (Table 1). A seeded SHA-256 hash of the market id chooses C1’s shared pair, and C3 assigns two members to each of (A_1, A_2) , (A_1, A_3) , and (A_2, A_3) . C1 and C3 therefore match twelve article exposures across members exactly while increasing unique panel-level coverage from two articles to three. On the strict sample, the mean member receives 3,909 article-text characters in C1 and 3,918 in C3, so total panel exposure differs by only 0.23% (C3–C1: +53.4 characters, paired-bootstrap 95% CI [−14.8, +119.6]). Exposure counts are matched exactly;

this character audit does not claim exact tokenizer-level matching.

The collection also includes role-conditioned C2 and C4 panels. Because they do not isolate evidence routing among neutral members, we treat them as exploratory. Their design is documented in Appendix B, and their complete outputs remain in the released artifact.

Served endpoints and protocol. We ran GPT-4.1-mini (gpt-4.1-mini-2025-04-14), Qwen (Qwen/Qwen3-30B-A3B-Instruct-2507), and OLMo (allenai/olmo-3.1-32B-Instruct) on August 4 and 5, 2026. The open-weight models were served by vLLM 0.11.0 on one H100 in bfloat16. No Hugging Face revision was pinned prospectively in the launch command, but the persistent serving cache contained exactly one snapshot for each repository, so we recovered and report both resolved revisions.¹ Their cache timestamps match the corresponding run dates. The client requested temperature 0.7, top- p 0.95, and 512 maximum output tokens. Qwen additionally retained its model-provided top- k = 20 because the client did not set top- k . We define an endpoint as this full served model and decoding stack, not as an architecture family.

Calls had no tools, browsing, or live external retrieval and received no market prices or order-book state, so the assigned static article text was their only event evidence. Each independent call returns JSON with a probability, a short rationale, and article ids; schema-invalid outputs are retried once at temperature zero. Across all six collected conditions, there are 21,948 expected member records per endpoint: 21,948 parsed for GPT, 21,948 for OLMo, and 21,947 for Qwen because of one missing exploratory C2 output. The strict primary sub-

¹Qwen revision:
0d7cf23991f47feeb3a57ecb4c9cee8ea4a17bfe
OLMo revision:
ac0587e4a7744a551c059d8cd17ba220bc940dae.

	member	dispersion	ensemble
<i>GPT-4.1-mini</i>			
C0 single	n/a	n/a	0.2263
C1 shared pair	0.2306	0.0027	0.2280
C5 shared, full	0.2278	0.0033	0.2245
C3 routed pairs	0.2350	0.0119	0.2231
<i>Qwen3-30B-A3B-Instruct</i>			
C0 single	n/a	n/a	0.2290
C1 shared pair	0.2337	0.0026	0.2311
C5 shared, full	0.2275	0.0025	0.2250
C3 routed pairs	0.2292	0.0091	0.2201
<i>OLMo-3.1-32B-Instruct</i>			
C0 single	n/a	n/a	0.2488
C1 shared pair	0.2391	0.0039	0.2353
C5 shared, full	0.2459	0.0040	0.2419
C3 routed pairs	0.2424	0.0113	0.2311

Table 2: Strict pre-end sample ($n = 609$). Ensemble equals member minus dispersion up to shown rounding. Lower member and ensemble Brier are better; higher dispersion is the contribution from arithmetic averaging.

set contains 18,879 complete member records per endpoint, and every C1 and C3 panel is complete. We use arithmetic means throughout, with no deliberation or supervisor stage.

Inference. We estimate paired per-market C3–C1 contrasts and unadjusted 95% percentile intervals from $B = 10,000$ bootstrap resamples. Let b_+ and b_- count bootstrap mean contrasts at or above and at or below zero, respectively, with ties counted in both tails, and use the plus-one Monte Carlo value $p = \min\{1, 2[\min(b_+, b_-) + 1]/(B + 1)\}$ so that finite simulation never yields an exact zero. Endpoint interactions are paired differences of differences on the same markets. We apply Holm’s step-down method separately to two declared families: the three endpoints by three channels (9 within-endpoint tests) and the three endpoint pairs by three channels (9 interaction tests). Other analyses are post-hoc sensitivities with unadjusted intervals. We apply separate Holm corrections to the six evidence-pair components and the nine Finance-tagged effects. Significance on one endpoint but not another is not an endpoint interaction.

5 Results

5.1 Primary effects

Routing raises dispersion by +0.00923 [+0.00762, +0.01098] for GPT, +0.00645 [+0.00534, +0.00766] for Qwen, and +0.00745 [+0.00608,

C3–C1, strict $n = 609$	Estimate [95% CI]
<i>Dispersion gained</i>	
GPT	+0.00923 [+0.00762, +0.01098]
Qwen	+0.00645 [+0.00534, +0.00766]
OLMo	+0.00745 [+0.00608, +0.00898]
GPT–Qwen int.	+0.00278 [+0.00115, +0.00448]
<i>Member Brier change</i>	
GPT	+0.00437 [−0.00345, +0.01210]
Qwen	−0.00456 [−0.01174, +0.00242]
OLMo	+0.00326 [−0.00365, +0.00997]
GPT–Qwen int.	+0.00893 [+0.00103, +0.01711]
<i>Ensemble Brier change</i>	
GPT	−0.00486 [−0.01263, +0.00276]
Qwen	−0.01101 [−0.01846, −0.00400]
OLMo	−0.00418 [−0.01115, +0.00256]
GPT–Qwen int.	+0.00616 [−0.00188, +0.01427]

Table 3: Primary contrast by channel with unadjusted 95% intervals. In the separate Holm families, Qwen’s ensemble effect has adjusted $p = 0.013$ among the 9 within-endpoint tests; only the dispersion difference between GPT and Qwen survives among the 9 endpoint interactions (adjusted $p = 0.005$).

+0.00898] for OLMo (Table 3). Every interval excludes zero in both the strict and full samples.

GPT’s increase exceeds Qwen’s by +0.00278 [+0.00115, +0.00448], with a Holm-adjusted p of 0.005. This is the only surviving interaction in the declared family. The differences between GPT and OLMo and between Qwen and OLMo do not survive correction, so this result does not rank model families.

Qwen’s ensemble Brier changes by −0.01101 [−0.01846, −0.00400], while GPT (−0.00486 [−0.01263, +0.00276]) and OLMo (−0.00418 [−0.01115, +0.00256]) remain inconclusive. Qwen’s Holm-adjusted p is 0.013 across the 9 within-endpoint tests (Table 8). The net difference between GPT and Qwen is +0.00616 [−0.00188, +0.01427] (Holm $p = 0.544$), so the net benefit is not established as different across endpoints.

Average member Brier rises descriptively for GPT (+0.00437) and OLMo (+0.00326) and falls for Qwen (−0.00456), but all three intervals include zero. The member-Brier difference between GPT and Qwen is +0.00893 [+0.00103, +0.01711] and does not survive correction (Holm $p = 0.240$).

5.2 Mechanism checks

Within C3, we split dispersion into variation between the three article-pair means and variation between repeated calls on each pair. C1 has no evidence groups, so we average the same

calculation over all 15 ways to partition its six calls into three pseudo-pairs. The between-pair component is $+0.0091$ [$+0.0075$, $+0.0108$] for GPT, $+0.0064$ [$+0.0054$, $+0.0076$] for Qwen, and $+0.0073$ [$+0.0060$, $+0.0088$] for OLMo. These values account for 98.2%, 99.9%, and 98.1% of the respective point-estimate dispersion increases. All three have Holm-adjusted $p = 0.0012$ in the six-test family. Every within-pair remainder interval includes zero (Appendix Table 10).

In a separate post-hoc sensitivity, 86 strict markets carry the overlapping platform tag *Finance*. Routing raises dispersion by $+0.0100$ [$+0.0065$, $+0.0138$] for GPT, $+0.0057$ [$+0.0038$, $+0.0080$] for Qwen, and $+0.0047$ [$+0.0026$, $+0.0068$] for OLMo. Their Holm-adjusted p values are 0.0018, 0.0018, and 0.0028. Every average-member-Brier and ensemble interval in this subset includes zero (Appendix Table 10).

The primary $C1 \rightarrow C3$ effect combines an expansion of the panel’s evidence union with a more even pair allocation. From the frozen calls, we construct a concentrated $P_{4:1:1}$ panel in which four forecasts use $C1$ ’s shared pair and one uses each other pair. It contains both $C3$ forecasts on the shared pair, two selected $C1$ forecasts on that pair, and one selected $C3$ forecast from each other pair, and we average each market’s decomposition over all $\binom{6}{2} \times 2 \times 2 = 60$ admissible selections. $P_{4:1:1}$ and $C3$ therefore both have six members, two articles per member, twelve exposures, and the same three-article union. Because neutral $C1$ and $C3$ calls receiving the shared pair use identical prompts and decoding requests, the construction treats them as exchangeable apart from sampling randomness.

Expanding $C1$ to $P_{4:1:1}$ increases dispersion and improves ensemble Brier on all three endpoints in this unadjusted sensitivity (Table 4). The coverage-expansion step accounts for most of the primary net point estimate at each endpoint, while balancing $P_{4:1:1}$ to $C3$ increases dispersion on every endpoint while holding the union fixed. Only Qwen’s balancing-step ensemble interval excludes zero; every balancing-step member interval includes zero. This post-hoc mixed panel is a mechanism check, not a separately randomized condition.

5.3 Practical and temporal checks

$C3$ ’s ensemble Brier is descriptively lower than a single full-evidence $C0$ call for all three endpoints, with the Qwen and OLMo intervals excluding zero. Against six full-evidence $C5$ calls, only OLMo’s

interval excludes zero. These secondary comparisons address whether routing justifies extra calls, but they cannot isolate routing itself, since $C0$ and $C5$ expose different total article text.

As external financial context, the strict sample has a 45.2% YES rate and climatology Brier 0.2477. On the 608 rows with a parseable contemporaneous market probability, that probability scores 0.1675, versus 0.2199 for the best routed model ensemble (Qwen) on the same rows. This study therefore diagnoses model routing; it is not evidence of outperforming the market.

Requiring snapshots to precede market end by more than 12 hours, one day, three days, or seven days leaves every endpoint’s dispersion interval above zero (Table 6). Qwen’s ensemble interval stays below zero through the three-day cutoff; at seven days its estimate is still favorable, but the interval crosses zero as the sample falls to 123. This audit cannot date individual articles, but the pattern is not confined to snapshots taken immediately before market end.

6 Readout sensitivity

Qwen verbalizes only 27 distinct probability values in the readout dataset, and 47.3% of $C1$ markets yield six identical reported numbers. For each completed Qwen $C0$, $C1$, and $C3$ forecast, we reconstructed the original neutral system turn and user turn (question, resolution criteria, and assigned article ids, metadata, and text), then replayed only the stored rationale as the assistant turn. We did not replay the stored `probability_yes` or `predicted_outcome` values, though numeric language already present in the rationale was retained. No new rationale was generated.

We appended a forced choice between YES and NO as A and B in both label orders. At temperature zero and one maximum output token, the server generated one answer token and returned the top 20 alternatives at that first position, and all 18,408 emitted tokens decoded to A or B. We scored single-token alternatives that decoded to A or B after whitespace stripping and case normalization. If one label was absent from the top 20, its log probability was set to -12 ; no balanced score was missing. We compute $s = \frac{1}{2}(d_1 - d_2)$, where $d = \log P(A) - \log P(B)$, to cancel label and position preference to first order. The score has 7,578 distinct stored values (729 after rounding to six decimals) and reduces identical $C1$ panels to 0.0%. We

Endpoint	Step	Δ member [95% CI]	Δ dispersion [95% CI]	Δ ensemble [95% CI]
GPT	$P_{4:1:1} - C1$	+0.00196 [−0.00219, +0.00603]	+0.00686 [+0.00564, +0.00825]	−0.00491 [−0.00918, −0.00086]
	$C3 - P_{4:1:1}$	+0.00242 [−0.00135, +0.00619]	+0.00237 [+0.00185, +0.00292]	+0.00005 [−0.00362, +0.00373]
Qwen	$P_{4:1:1} - C1$	−0.00232 [−0.00612, +0.00136]	+0.00495 [+0.00406, +0.00598]	−0.00727 [−0.01139, −0.00344]
	$C3 - P_{4:1:1}$	−0.00224 [−0.00573, +0.00119]	+0.00150 [+0.00113, +0.00189]	−0.00374 [−0.00719, −0.00042]
OLMo	$P_{4:1:1} - C1$	+0.00165 [−0.00198, +0.00527]	+0.00535 [+0.00442, +0.00640]	−0.00369 [−0.00746, −0.00008]
	$C3 - P_{4:1:1}$	+0.00161 [−0.00177, +0.00494]	+0.00210 [+0.00156, +0.00268]	−0.00049 [−0.00387, +0.00286]

Endpoint	C3–C0 ensemble [95% CI]	C3–C5 ensemble [95% CI]
GPT	−0.00320 [−0.01120, +0.00463]	−0.00133 [−0.00821, +0.00538]
Qwen	−0.00897 [−0.01668, −0.00146]	−0.00490 [−0.01063, +0.00066]
OLMo	−0.01773 [−0.02655, −0.00925]	−0.01087 [−0.01812, −0.00409]

Table 4: Post-hoc unadjusted strict-sample checks. Top: $C1 \rightarrow P_{4:1:1}$ expands the union; $P_{4:1:1} \rightarrow C3$ balances routing at a fixed union. Bottom: practical net comparisons. C0 is one full-evidence call; C5 is six full-evidence calls, so neither comparison matches C3’s call/input budget.

Endpoint	Sample	Δ member [95% CI]	Δ dispersion [95% CI]	Δ ensemble [95% CI]
GPT	strict	+0.00437 [−0.00345, +0.01210]	+0.00923 [+0.00762, +0.01098]	−0.00486 [−0.01263, +0.00276]
GPT	full	+0.00441 [−0.00293, +0.01175]	+0.00934 [+0.00783, +0.01102]	−0.00494 [−0.01225, +0.00223]
Qwen	strict	−0.00456 [−0.01174, +0.00242]	+0.00645 [+0.00534, +0.00766]	−0.01101 [−0.01846, −0.00400]
Qwen	full	−0.00427 [−0.01086, +0.00228]	+0.00692 [+0.00583, +0.00809]	−0.01119 [−0.01805, −0.00461]
OLMo	strict	+0.00326 [−0.00365, +0.00997]	+0.00745 [+0.00608, +0.00898]	−0.00418 [−0.01115, +0.00256]
OLMo	full	+0.00022 [−0.00655, +0.00701]	+0.00770 [+0.00634, +0.00918]	−0.00748 [−0.01439, −0.00071]

Table 5: Strict-versus-full temporal-filter sensitivity. Dispersion is stable. GPT and Qwen net conclusions match the strict analysis; OLMo’s full-sample ensemble interval excludes zero, whereas its stricter pre-end interval does not. This difference motivates the temporal filter rather than an OLMo gain claim.

call this the rationale-conditioned balanced-logit readout.

Because s is not intrinsically a probability, we map it to $[0, 1]$ with one logistic calibrator pooled across C0/C1/C3 and cross-fitted in five folds by market. For comparability, the verbalized probability is passed through the same class of pooled, cross-fitted map. The strict analysis fits and evaluates both maps using only the 609 strict markets, and a separate full-sample sensitivity fits and evaluates them on all 708. The control therefore compares two calibrated readouts, and its scale is not the raw Qwen scale used in Table 3. Both fold assignments are released.

From verbalized to logit readout, the ensemble contrast changes by +0.00047 [−0.00492, +0.00595], the dispersion component by +0.00310 [+0.00232, +0.00389], and the member component by +0.00356 [−0.00180, +0.00905] (Figure 1). The net change is not statistically detectable, but the allocation across channels differs. Equivalent stored-logit controls are unavailable for GPT and OLMo.

On all 708 markets, the corresponding logit-minus-verbalized interaction is +0.00399 [−0.00088, +0.00895] for member Brier, +0.00291 [+0.00218, +0.00364] for dispersion, and +0.00108 [−0.00381, +0.00603] for ensemble

Brier. The qualitative conclusion is unchanged: channel attribution moves while the net interaction remains statistically inconclusive.

7 Discussion and conclusion

Evidence routing increases dispersion in these parallel ensembles, but that does not guarantee a better forecast. The exact decomposition separates the contribution of dispersion from the change in average member Brier. The post-hoc evidence-pair attribution analysis places nearly all of the measured increase between evidence-pair groups rather than between repeated calls on the same pair. Under arithmetic probability averaging and Brier evaluation, reporting all three terms should be routine practice whenever a forecasting system claims a gain from role, prompt, retrieval, or sampling diversity.

Qwen’s second readout produces no statistically detectable change in the net contrast, yet it changes the measured channel balance. The identity is exact for either readout; the terms change because the input probabilities change.

For deployment, dispersion and average member Brier should be evaluated together under the same readout and served stack. Across the three endpoints studied here, broader panel coverage and

Endpoint	Channel	> 12 h (n = 575)	> 1 d (n = 538)	> 3 d (n = 348)	> 7 d (n = 123)
GPT	member	+0.0043 [−0.0038, +0.0124]	+0.0031 [−0.0054, +0.0114]	+0.0024 [−0.0083, +0.0130]	+0.0073 [−0.0091, +0.0238]
	dispersion	+0.0090 [+0.0073, +0.0108]	+0.0089 [+0.0072, +0.0108]	+0.0091 [+0.0068, +0.0117]	+0.0103 [+0.0067, +0.0146]
	ensemble	−0.0047 [−0.0129, +0.0031]	−0.0058 [−0.0142, +0.0021]	−0.0067 [−0.0174, +0.0035]	−0.0030 [−0.0192, +0.0125]
Qwen	member	−0.0030 [−0.0101, +0.0040]	−0.0042 [−0.0117, +0.0032]	−0.0051 [−0.0148, +0.0047]	−0.0016 [−0.0203, +0.0168]
	dispersion	+0.0063 [+0.0051, +0.0075]	+0.0063 [+0.0051, +0.0077]	+0.0064 [+0.0049, +0.0080]	+0.0075 [+0.0048, +0.0105]
	ensemble	−0.0093 [−0.0165, −0.0023]	−0.0105 [−0.0183, −0.0032]	−0.0115 [−0.0212, −0.0021]	−0.0091 [−0.0275, +0.0083]
OLMo	member	+0.0029 [−0.0040, +0.0099]	+0.0019 [−0.0052, +0.0090]	+0.0045 [−0.0045, +0.0134]	+0.0042 [−0.0117, +0.0201]
	dispersion	+0.0074 [+0.0060, +0.0089]	+0.0071 [+0.0057, +0.0086]	+0.0075 [+0.0055, +0.0096]	+0.0080 [+0.0050, +0.0117]
	ensemble	−0.0045 [−0.0116, +0.0026]	−0.0051 [−0.0123, +0.0020]	−0.0030 [−0.0123, +0.0062]	−0.0038 [−0.0195, +0.0113]

Table 6: C3–C1 effects [unadjusted 95% CI] after increasingly strict snapshot-to-market-end lead-time requirements.

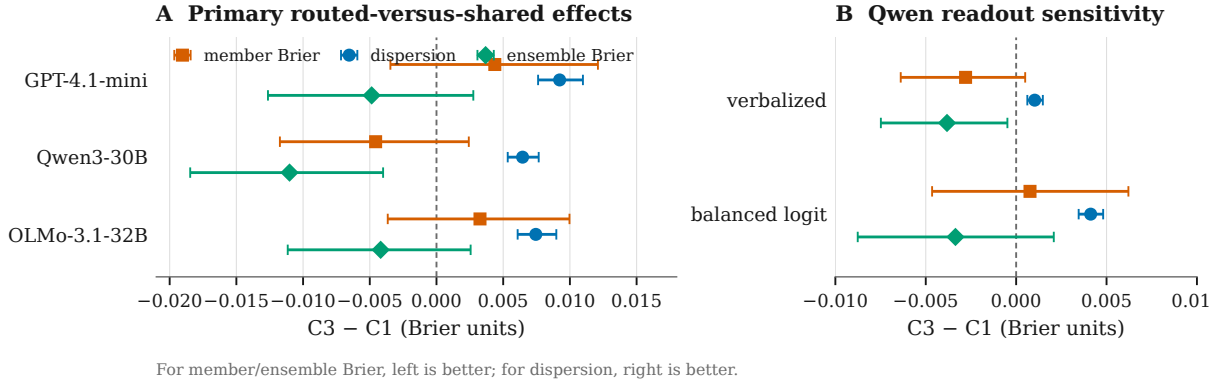


Figure 1: Paired bootstrap estimates and 95% intervals on the strict sample. **A**: raw verbalized C3–C1 effects for all three endpoints. Positive dispersion is beneficial; negative member and ensemble Brier are beneficial. **B**: Qwen after identical cross-fitted calibration of verbalized and balanced-logit readouts. The readout-induced change in the net contrast is not statistically detectable, while the channel decompositions differ.

Readout	Δ member	Δ dispersion	Δ ensemble
verbalized	−0.00280	+0.00103	−0.00382
balanced logit	+0.00077	+0.00412	−0.00336

Table 7: Qwen C3–C1 on the strict sample after identical cross-fitted calibration. The corresponding 95% intervals are: verbalized member [−0.00638, +0.00050], dispersion [+0.00062, +0.00148], ensemble [−0.00748, −0.00048]; logit member [−0.00465, +0.00621], dispersion [+0.00346, +0.00481], ensemble [−0.00876, +0.00208].

the post-hoc fixed-union balancing check both raise dispersion. The data establish neither a general endpoint ranking nor a common member-Brier cost. The Finance-tagged subset preserves the dispersion result but not a detectable net gain.

Limitations

The study is exploratory and was not preregistered. Data collection expanded after an initial 300-market batch appeared underpowered, yielding the 609 pre-end markets used for primary inference. The readout control was specified after inspection

of the primary results. Conditions, prompts, and served model identifiers were fixed before each corresponding run, but the study as a whole is adaptive, and the reported intervals do not adjust for optional stopping.

Temporal auditing is incomplete. The strict filter removes snapshots at or after market end, and lead-time effects persist well before end, but article-level dates are unavailable. The full-sample OLMo result shows why that distinction matters. Frozen, overlapping platform tags show that 68.5% of strict markets carry *Sports* and 14.1% carry *Finance*; they are event metadata rather than an exclusive taxonomy, so the Finance-tagged sensitivity remains exploratory. Text-only forecasting is not a trading evaluation; prices and order books are deliberately excluded from model inputs.

C1 and C3 match article counts and exposures, not the union of unique evidence, and that broader union is part of the primary treatment. The post-hoc $P_{4:1:1}$ control holds the union fixed but mixes calls collected under C1 and C3 rather than prospectively randomizing a separate condition. The evidence-pair attribution analysis likewise uses

actual C3 groups but averages C1 over post-hoc pseudo-pairings rather than a prospectively randomized grouping. Article-text characters are nearly matched on average, but exact tokens are not, and articles are truncated at 2,000 characters. A future design should randomize routing while holding both the evidence union and the tokenizer-level input budget fixed.

Only Qwen has a second readout. Its calibrators are cross-fitted but not refit inside each bootstrap resample, so their intervals omit calibration-fit uncertainty. Although the stored structured probability was omitted from the readout replay, the rationale was not scrubbed of numerical language, so rationale-mediated anchoring remains possible. The retained readout cache stores only the difference between A and B, not per-label top-20 membership, so the frequency of invoking the -12 fallback and a post-hoc floor sweep cannot be reconstructed without new inference. We ran only three served stacks at one temperature, and Qwen retained an uncontrolled model-default top- k . The open-weight revisions were recovered retrospectively from unique serving-cache snapshots rather than pinned prospectively in the launch command. These facts limit how far the results transfer beyond the recorded endpoints.

Data and Code Availability

Forecasts, assignments, metadata, prompts, and analysis code are available in the [supplementary artifact](#) and [public code repository](#). Because records store parsed forecasts rather than verbatim provider payloads, exact raw-response reconstruction is not possible.

Ethics Statement

The study uses public market and news data and makes no decisions about individuals. Forecasting systems may encourage financially consequential overconfidence; these diagnostic results are not investment advice or evidence that the models outperform market prices.

Acknowledgments

We thank Modal for providing compute credits.

References

- Matthew Aitchison, Scott Jeen, Toby Shevlane, and Ben Day. 2026. Diversity is the Strength of the AI Crowd. *arXiv:2606.29661*.
- Ege Beyazit, KL Navaneet, Prashant Mathur, Roi Blanco, Vidit Bansal, and Karim Bouyarmane. 2025. Enabling Fine-Grained Operating Points for Black-Box LLMs. *arXiv:2510.17727*.
- Pu Cheng, Juncheng Liu, and Yunshen Long. 2026. PolyBench: Benchmarking LLM Forecasting and Trading Capabilities on Live Prediction Market Data. *arXiv:2604.14199*.
- Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching Human-Level Forecasting with Language Models. *arXiv:2402.18563*.
- Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. 2024. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. *arXiv:2409.19839*.
- Hankyeol Kim and Pilsung Kang. 2026. Asking Is Not Enough: Protocol Sensitivity in LLM Confidence Calibration. *arXiv:2605.27752*.
- Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems* 7.
- Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. 2026. Diverse Evidence, Better Forecasts: Multi-Agent Deliberation Under Information Asymmetry. *arXiv:2607.01661*.
- Team OLMo. 2025. Olmo-3.1-32B-Instruct model card. *Hugging Face model documentation*. <https://huggingface.co/allenai/Olmo-3.1-32B-Instruct>.
- Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. 2024. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. *arXiv:2402.19379*.

A Complete statistical tables

Endpoint	Channel	Estimate	95% CI	Raw p	Holm p
GPT	member	+0.00437	[−0.00345, +0.01210]	0.2744	1.000
GPT	dispersion	+0.00923	[+0.00762, +0.01098]	0.0002	0.002
GPT	ensemble	−0.00486	[−0.01263, +0.00276]	0.2080	1.000
Qwen	member	−0.00456	[−0.01174, +0.00242]	0.2032	1.000
Qwen	dispersion	+0.00645	[+0.00534, +0.00766]	0.0002	0.002
Qwen	ensemble	−0.01101	[−0.01846, −0.00400]	0.0022	0.013
OLMo	member	+0.00326	[−0.00365, +0.00997]	0.3484	1.000
OLMo	dispersion	+0.00745	[+0.00608, +0.00898]	0.0002	0.002
OLMo	ensemble	−0.00418	[−0.01115, +0.00256]	0.2354	1.000

Table 8: Complete declared within-endpoint family on the strict sample. Each estimate is C3–C1. Intervals and raw p -values are unadjusted; Holm p -values correct jointly over all nine rows. Lower member and ensemble Brier and higher dispersion are favorable.

Channel	Endpoint pair (first–second)	Estimate	95% CI	Raw p	Holm p
member	GPT–Qwen	+0.00893	[+0.00103, +0.01711]	0.0300	0.240
member	GPT–OLMo	+0.00111	[−0.00747, +0.00960]	0.8107	1.000
member	Qwen–OLMo	−0.00782	[−0.01616, +0.00033]	0.0590	0.354
dispersion	GPT–Qwen	+0.00278	[+0.00115, +0.00448]	0.0006	0.005
dispersion	GPT–OLMo	+0.00178	[+0.00001, +0.00358]	0.0490	0.343
dispersion	Qwen–OLMo	−0.00099	[−0.00263, +0.00058]	0.2130	0.639
ensemble	GPT–Qwen	+0.00616	[−0.00188, +0.01427]	0.1280	0.544
ensemble	GPT–OLMo	−0.00067	[−0.00907, +0.00777]	0.8669	1.000
ensemble	Qwen–OLMo	−0.00683	[−0.01531, +0.00155]	0.1088	0.544

Table 9: Complete declared endpoint-interaction family on the strict sample. Estimates are differences between endpoint-specific C3–C1 effects in the listed direction. Intervals and raw p -values are unadjusted; Holm p -values correct jointly over all nine rows.

Endpoint	Component	Δ dispersion [95% CI]	Holm p
GPT	between pair	+0.0091 [+0.0075, +0.0108]	0.0012
	within pair	+0.0002 [−0.0001, +0.0005]	0.7433
Qwen	between pair	+0.0064 [+0.0054, +0.0076]	0.0012
	within pair	+0.0000 [−0.0003, +0.0003]	0.9965
OLMo	between pair	+0.0073 [+0.0060, +0.0088]	0.0012
	within pair	+0.0001 [−0.0001, +0.0004]	0.7433

Endpoint	Channel	C3–C1 [95% CI]	Holm p
GPT	member	+0.0022 [−0.0172, +0.0207]	1.0000
	dispersion	+0.0100 [+0.0065, +0.0138]	0.0018
	ensemble	−0.0078 [−0.0271, +0.0102]	1.0000
Qwen	member	−0.0026 [−0.0163, +0.0111]	1.0000
	dispersion	+0.0057 [+0.0038, +0.0080]	0.0018
	ensemble	−0.0083 [−0.0219, +0.0050]	1.0000
OLMo	member	+0.0049 [−0.0103, +0.0199]	1.0000
	dispersion	+0.0047 [+0.0026, +0.0068]	0.0028
	ensemble	+0.0003 [−0.0151, +0.0152]	1.0000

Table 10: Post-hoc strict-sample analyses. Left: evidence-pair attribution analysis; the two components sum to the primary dispersion contrast, and Holm correction is over six rows. Right: the 86 markets carrying the overlapping platform tag *Finance*; Holm correction is over nine rows. All intervals are unadjusted.

B Exploratory role-conditioned conditions

The auxiliary C2 condition applies the same six role prompts to members sharing one article pair; C4 uses those roles with two members per pair and permutes the role-to-pair map by market. These conditions do not isolate evidence routing among neutral members, so we make no role-prompting inference. Their complete forecasts are included in the released artifact for transparency.